

***Combined in silico approaches towards the identification of
novel malarial cysteine protease inhibitors***

A thesis submitted in fulfilment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY
IN BIOINFORMATICS

of

RHODES UNIVERSITY, SOUTH AFRICA

Research Unit in Bioinformatics (RUBi)

DEPARTMENT OF BIOCHEMISTRY and MICROBIOLOGY

Faculty of Science

by

THOMMAS MUTEMI MUSYOKA

February 2016



RHODES UNIVERSITY
Where leaders learn



ABSTRACT

Malaria an infectious disease caused by a group of parasitic organisms of the *Plasmodium* genus remains a severe public health problem in Africa, South America and parts of Asia. The leading causes for the persistence of malaria are the emergence of drug resistance to common antimalarial drugs, lack of effective vaccines and the inadequate control of mosquito vectors. Worryingly, accumulating evidence shows that the parasite has developed resistant to the current first-line treatment based on artemisinin. Hence, the identification and characterization of novel drug targets and drugs with unique mode of action remains an urgent priority. The successful sequencing and assembly of genomes from several *Plasmodium* species has opened an opportune window for the identification of new drug targets. Cysteine proteases are one of the major drug targets to be identified so far. The use of cysteine protease inhibitors coupled with gene manipulation studies has defined specific and putative roles of cysteine proteases which include hemoglobin degradation, erythrocyte rupture, immune evasion and erythrocyte invasion, steps which are central for the completion of the *Plasmodium* parasite life cycle.

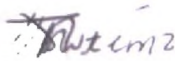
In an aim to discover potential novel antimalarials, this thesis focussed on falcipains (FPs), a group of four papain-like cysteine proteases from *Plasmodium falciparum*. Two of these enzymes, FP-2 and FP-3 are the major hemoglobinases and have been validated as drug targets. For the successful elimination of malaria, drugs must be safe and target both human and wild *Plasmodium* infective forms. Thus, an incipient aim was to identify protein homologs of these two proteases from other *Plasmodium* species and the host (human). From BLASTP analysis, up to 16 FP-2 and FP-3 homologs were identified (13 *plasmodial* proteases and 3 human cathepsins). Using *in silico* characterization approaches, the intra and inter group sequence, structural, phylogenetic and physicochemical differences were determined. To extend previous work (MSc student) involving docking studies on the

identified proteins using known FP-2 and FP-3 inhibitors, a South African natural compound and its ZINC analogs, molecular dynamics and binding free energy studies were performed to determine the stabilities and quantification of the strength of interactions between the different protein-ligand complexes. From the results, key structural elements that regulate the binding and selectivity of non-peptidic compounds onto the different proteins were deciphered. Interaction fingerprints and energy decomposition analysis identified key residues and energetic terms that are central for effective ligand binding.

This research presents novel insight essential for the structure-based molecular drug design of more potent antimalarial drugs.

DECLARATION

I, **THOMMAS MUTEMI MUSYOKA**, declare that this is my own unaided work, except where duly acknowledged. It is being submitted for the degree of Doctor of Philosophy in Bioinformatics in the Faculty of Science of Rhodes University. It has not been submitted before for any degree for examination in any other university.



THOMMAS MUTEMI MUSYOKA

DATED THIS 25th **DAY OF** AUGUST **2016**

DEDICATION

To my dearest dad, family, mentors and friends without whom I could hardly complete my thesis. You are all my heroes and heroines.

ACKNOWLEDGEMENTS

For the successful completion of this work, many people from different backgrounds, statuses and disciplines have contributed in one way or the other. To you all, I owe my special gratitude and blessings for without your input I would have given up long ago.

My deepest gratitude to my main supervisor Prof. Özlem Tastan Bishop for her endless contributions in terms of time, insights and funds that ensured my PhD life was productive and stimulating. My interactions with her have not only taught me about bioinformatics, but other success principles which I can identify with the words “*where there is will there is a way*”. I believe this will always fuel my determination in the pursuit of award winning accomplishments in my future endeavours. I have been amazingly fortunate to have her as my mentor. Her patience and enthusiasm has transformed a greenhorn in programming to someone who can communicate effectively with computers. Her limitless kindness, care, patience, unwavering support and friendship have made my three years that I have been at RUBi seem to be like a period of few months.

To my co-supervisor, Dr. Kevin Lobb, I lack words that can fully explain my gratitude for the role he played in mentoring me. He consciously and unconsciously taught me how I can be better in the field of computational chemistry. His calmness and willingness to pull me through the computational chemistry side of things will forever be treasured. When I first met him in 2013 and he introduced me to the data that was to form the basis of my PhD project, I felt like giving up considering my neophytism in computational programming and modelling. However, he patiently mentored me all through and what seemed to be beyond my capacity has turned out to a thing I enjoy doing. His dedication, financial support and concern for my wellbeing for the last three years will be cherished forever.

I am also greatly thankful to the RUBi team, its past and present team whom I have interacted with during the past three years. Your moral support and the many jokes we shared made me stay sane in a foreign land far from my home. The famous quote we coined “*never leave a fellow soldier in the woods*” has really found a place in my life. You will forever remain inscribed in my heart and will cherish your friendship evermore.

To my many friends that I have met along the way, you have all in a way supported my quest for this lifetime achievement. As I cannot afford to mention you all, I hope my gratefulness for your moral and material support will find a place in your hearts. I am greatly indebted for the support from Caleb K. Kipkurui, Hafeni Mthoko and Rosaline Macharia. Special thanks to Mr and Mrs David Mbithi, Mr and Mrs Josphat Tama, Mr and Mrs John Gatimu, Mr and Mrs Raphael Kingola, Dr. Chris Vita and his family, Mr and Mrs Stephen Museveni and the family of Mr. Jechoniah Kimolo kitala.

I am greatly indebted to the Redeemed Christian Church of God – House of Praise brethren for their support and prayers. Though I was far away from home, you played exceptionally well the role of brothers and sisters in my life. My sister Dr. Vanessa and my brother Dr. Komlan, thank you for your support and love.

To my dearest dad, brothers and sisters, your prayers have always kept me strong. Your constant love, concern and constant messages reminding me that it is possible have motivated me through even at the darkest valley of my journey. To my fiancée, Bridget K. Mutuma, your belief in me and love has motivated me all through. I will forever treasure you.

Finally, I thank God for your unfailing grace, mercies and faithfulness. In deed your scripture in Psalms 23: 1- 6 has become my testimony. I will forever remain your obedient child.

Funding acknowledgement

The achievements so far made would not have been possible without the generous financial support from the Rhodes University research office. Kudos to you all for making the realization of my dreams. I am greatly indebted to Rhodes University administration for awarding me with the Rhodes University Prestigious Postgraduate Scholarship that ensured I remained focused on my studies. This work was partially supported by the National Institutes of Health Common Fund (Grant number: U41HG006941 to H3ABioNet).

TABLE OF CONTENTS

ABSTRACT.....	i
DECLARATION.....	iii
DEDICATION.....	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vii
LIST OF FIGURES	xii
LIST OF TABLES	xvi
RESEARCH OUTPUTS	xvii
LIST OF ABBREVIATIONS	xx
THESIS OVERVIEW	xxii
CHAPTER 1	1
Malaria	1
1.1 Malaria: The ever elusive global health challenge	2
1.2 <i>Plasmodium</i> life cycle.....	3
1.3 Antimalarial drug history	4
1.3.1 The role of natural products in antimalarial drug discovery.....	6
1.3.2 Other antimalarial natural products	10
1.3.3 Synthetic antimalarial drugs	11
1.3.3.1 Classification of available antimalarial drugs.....	11
1.4 Antimalarial drug resistance	12
1.5 Antimalarial drug development in the post genomic era	13
1.5.1 Establishing a drug target in antimalarial drug development	14
1.5.2 Antimalarial drug targets	14
1.5.3 Drug development pipeline and the place of modern computer technology	17
1.5.4 The current status of the antimalarial drug pipeline	18
1.6 Problem statement and justification.....	19
1.7 Hypothesis.....	21
1.8 Research aim	21
CHAPTER 2	23
<i>Plasmodial</i> cysteine proteases: <i>In silico</i> characterization	23
2.1 Proteases	24
2.1.1 Cysteine protease nomenclature	25
2.2 Roles of <i>plasmodial</i> cysteine proteases	26

2.2.1 Haemoglobin hydrolysis	27
2.2.2 Tissue and erythrocyte invasion.....	28
2.2.3 Erythrocyte rupture	29
2.2.4 Immunoavoidance.....	29
2.2.5 Exo-erythrocytic parasite stages	29
2.3 FPs cysteine proteases.....	30
2.3.1 FPs expression profiles in iRBCs	30
2.3.2 Biochemical characterization of FPs	31
2.3.3 The structure and functions of the different FPs domains	31
2.3.3.1 Falcipain prodomain	32
2.3.3.2 Mature domain	33
2.4 Structural basis of falcipain inhibition.....	33
2.5 Proposed work	34
2.6 Methodology	35
2.6.1 Data acquisition	35
2.6.1.1 Protein amino acid sequence retrieval	35
2.6.1.2 Retrieval of 3D protein structures.....	35
2.6.2 Multiple sequence alignment and subsite composition analysis.....	36
2.6.3 Phylogenetic analysis.....	36
2.6.4 Motif discovery	37
2.6.5 Physicochemical properties	37
2.7 Results and Discussion	38
2.7.1 Sequence analysis	38
2.7.1.1 Protein sequences.....	38
2.7.1.2 Multiple sequence analysis	39
2.7.1.3 Structure and composition of the binding pocket of cysteine proteases.....	41
2.7.2 Phylogenetic analysis.....	43
2.7.3 Motif analysis.....	45
2.7.4 Physicochemical properties	48
2.8 Chapter conclusion.....	50
CHAPTER 3	51
Molecular Dynamics Simulation Studies	51
3.1 Introduction.....	52
3.1.2 MD limitations	53
3.1.3 MD simulations history in chemistry and biology.....	54

3.1.4 MD dynamics in drug design.....	57
3.2 Conventional MD simulations.....	58
3.2.1 A model of the system.....	58
3.2.2 Force fields.....	59
3.2.3 Integration of Newtonian equation of motion.....	61
3.2.4 Solvation models.....	62
3.3 Proposed work.....	63
3.4 Methodology.....	66
3.4.1 Preparation of Protein-Ligand Complexes.....	67
3.4.2 MD Simulation.....	67
3.4.2.1 System set up.....	67
3.4.2.2 Preparation of protein and ligand topology files.....	68
3.4.2.3 Explicit solvent simulation parameters.....	69
3.4.3 Post-dynamic analysis.....	70
3.4.4 System specifications.....	71
3.4.5 Drug-likeness of identified hits.....	71
3.4.6 MD pipeline.....	72
3.4.6.1 Initial preparation.....	72
3.4.6.2 Force field conversion.....	73
3.4.6.3 Solvation.....	74
3.4.6.4 Neutralization.....	74
3.4.6.5 Energy minimization.....	74
3.4.6.6 Equilibration.....	76
3.4.6.7 Final production run.....	76
3.5 Results and Discussion.....	77
3.5.1 Quality assurance.....	78
3.5.2 Visualization.....	79
3.5.3 CPs.....	80
3.5.3.1 RMSD.....	81
3.5.3.2 Rg.....	83
3.5.3.3 RMSF.....	85
3.5.3.4 Solvent Accessible Surface Area (SASA).....	87
3.5.3.5 Binding mode.....	88
3.5.3.6 Structural chemical features of binding.....	89
3.5.3.7 Conformational changes during simulation.....	90

3.5.4 5PGA and selected ZINC hits.....	92
3.5.4.2 RMSD	92
3.5.4.3 Rg.....	95
3.5.4.4 RMSF	96
3.5.4.5 Hydrogen bonding	96
3.5.4.6 Secondary structure element stability	100
3.5.4.7 Binding mode.....	101
3.5.5 Chemical modifications necessary for hit to lead compounds.....	103
3.5.6 MD pipeline	104
3.6 Chapter conclusion.....	106
CHAPTER 4.....	108
Binding free energy calculations.....	108
4.1 Protein molecular recognition.....	109
4.2 Computational methods for BFE determination	110
4.2.1 Empirical methods for BFE calculations	110
4.2.2 Molecular force field methods for BFE calculations.....	111
4.3 Motivation.....	111
4.4 Methodology	113
4.4.1 Preparation of input files.....	113
4.4.2 Executing g_mmpbsa.....	113
4.4.3 Analysis.....	114
4.4.4 System specifications.....	115
4.5 Results and discussion	115
4.5.1 Role of BFE terms in the ligand binding process	115
4.5.2 Subsites contribution to BFE	118
4.5.3 Structural features affecting BFE.....	120
4.6 Chapter conclusion.....	122
CHAPTER 5.....	123
Docking studies: New <i>plasmodial</i> cysteine inhibitors from SANCDB.....	123
5.1 Introduction.....	124
5.2 South African Natural Compound Database (SANCDB).....	125
5.3 Docking software	126
5.4 AutoDock	127
5.5 Docking types	128
5.6 Methodology	129

5.6.1 Protein structure data and ligands	129
5.6.2 Docking studies.....	130
5.6.2.1 Preparation of protein and ligand files.....	130
5.6.2.2 Grid evaluation and affinity maps determinations.....	130
5.6.3 Docking simulation	131
5.6.4 Analysis.....	131
5.6.5 System specifications.....	132
5.7 Results and Discussion	132
5.7.1 Identification of best hits	132
5.7.2 Molecular interactions of best hits.....	133
5.7.3 Comparison with CPs, 5PGA and ZINC hits.....	136
5.8 Chapter conclusion.....	137
CHAPTER 6	138
Conclusions and future prospects	138
6.1 Conclusions.....	139
6.2 Future prospects	141
References	143

LIST OF FIGURES

Figure 1.1: <i>Plasmodium</i> parasite life cycle.....	3
Figure 1.2: 2D chemical structures of the early antimalarial drugs.....	7
Figure 1.3: Artemisinin and its derivatives which are the basis of current antimalarial therapies.....	9
Figure 1.4: Modern drug discovery pipeline	18
Figure 2.1: Catalytic mechanism of cysteine proteases	25
Figure 2.2: Nomenclature of cysteine (thiol) proteases.....	26
Figure 2.3: The role of cysteine proteases in the <i>Plasmodium</i> erythrocytic stage.....	27
Figure 2.4: Papain proteases structure	32
Figure 2.5: Analytic approaches applied to FP-2, FP-3 and homologs	35
Figure 2.6: Sequence analysis.....	40
Figure 2.7: The structure of the cathepsins and FPs.....	43
Figure 2.8: Evolution analysis	44
Figure 2.9: Motif analysis.....	46
Figure 2.10: Location of motifs	47
Figure 2.11: Protein composition analysis.....	49
Figure 3.1: Biophysical techniques and their applications	53
Figure 3.2: The time evolution of key developments in MD simulation and the resulting effects in simulation length of BPTI.....	55
Figure 3.3: Commonly used MD software	56
Figure 3.4: A schematic view of force field interactions in a molecular system.....	59
Figure 3.5: 2D structures of known FP-2 and/or FP-3 non-peptidic compounds.....	63
Figure 3.6: 5 α -Pregna-1,20-dien-3-one and its analogues from the ZINC database.....	65
Figure 3.7: MD simulation overview.....	66

Figure 3.8: The ligand separator interface which acts as the first stage in the MD simulation process.....	72
Figure 3.9: The force field conversion interface.....	73
Figure 3.10: The solvation interface.....	74
Figure 3.11: Genion tool interface.....	75
Figure 3.12: The energy minimization tool.....	75
Figure 3.13: Temperature equilibration tool.....	76
Figure 3.14: The isobaric equilibration tool.....	76
Figure 3.15: Production tool interface.....	77
Figure 3.16: Kinetic energy of FP-2, FP-3 and human cathepsins.....	79
Figure 3.17: A triclinic simulation box.....	80
Figure 3.18: Global conformational diversity of <i>plasmodial</i> and human proteases when in complex with CPs.....	82
Figure 3.19: Trajectory plots showing RMSD fluctuations.....	83
Figure 3.20: Ligand RMSD fluctuations.....	84
Figure 3.21: Compactness of the different protein-ligand complexes.....	84
Figure 3.22: Local residue fluctuations.....	86
Figure 3.23: The average local fluctuations of the subsite residues.....	87
Figure 3.24: The docking pose of CPG (blue), CPH (magenta) and CPI (green) in the binding pocket of human cathepsins and <i>plasmodial</i> proteases.....	88
Figure 3.25: The residues interacting with CPG (blue), CPH (magenta) and CPI (green) in the binding pocket of human cathepsins and <i>plasmodial</i> proteases during the docking stage.....	89
Figure 3.26: Hydrogen bond dynamic profiles.....	90
Figure 3.27: Ligand conformational changes over time.....	91
Figure 3.28: The RMSD plots of human cathepsins (Cat K and Cat L) and the FPs (FP-2 and FP-3) when in complex with a natural SA compound (5PGA) and selected analogs from the ZINC database.....	93

Figure 3.29: The average RMSD values of the apo (a) and holo (b) systems for the time period between 8 and 20 ns.....	94
Figure 3.30: The average RMSD values of 5PGA and its ZINC analogs for the time period between 8 and 20 ns.....	94
Figure 3.31: The average compactness of the different proteins when in complex with 5PGA and its analogs for the time period between 8 and 20 ns	95
Figure 3.32: Local residue fluctuations of falcipains (red and black) and cathepsins (green and blue) when complexed with 5PGA during the last 12 ns of a MD simulations.....	96
Figure 3.33: The number and evolution of intermolecular H-bonds	98
Figure 3.34: The evolution of H-bond length between Cat K Gln143 and ZINC03869631 and ligand orientation at different time points during MD simulation.....	99
Figure 3.35: The quantitative and qualitative analysis of H-bonds	99
Figure 3.36: Evolution of H-bond length between FP-2 Ile85 and ZINC03869631 and ligand orientation at different time points during MD simulation.....	100
Figure 3.37: Conformational evolution of secondary structure elements.....	101
Figure 3.38: Binding poses of 5PGA.....	102
Figure 3.39: Binding pocket aa residue interactions patterns.....	102
Figure 3.40: The automated MD simulation pipeline.....	105
Figure 3.41: A diagrammatic representation of how the different GROMACS tools for the MD pipeline tool are linked and the dependancies of each step.....	106
Figure 4.1: Box plots showing the distribution of the various BFE terms of CPs.....	117
Figure 4.2: Box plots showing the distribution of the various energy terms of the different proteins when bound to 5PGA and its ZINC analogs.....	118
Figure 4.3: A detailed per-residue fingerprint	119
Figure 4.4: Per-residue decomposition analysis	119
Figure 4.5: A detailed per-residue fingerprint	121

Figure 5.1: A diagrammatic representation of the different steps and tools used for docking studies	129
Figure 5.2: A stacked column chart showing the type and percentage of residues interacting with the best three hits identified	133
Figure 5.3: A surface presentation and corresponding binding pocket residue interaction network	135
Figure 5.4: A surface presentation and corresponding 2D interaction map	136

LIST OF TABLES

Table 1.1: Classification of known antimalarial drugs	12
Table 1.2: Key genes in <i>P. falciparum</i> and aa mutations leading to drug resistance	13
Table 1.3: Antimalarial drug candidates undergoing (pre-) clinical trials.....	19
Table 2.1: Key FP-2 and FP-3 homologs from different <i>plasmodial</i> proteases	38
Table 2.2: The position of the catalytic domain within the whole protein sequences of different FP-2 and FP-3 homologs.....	39
Table 2.3: A summary of the different physicochemical properties for FP-2, FP-3 and their homologs.....	48
Table 3.1: Key physicochemical properties for drug-like molecules	71
Table 3.2: The average of different thermodynamic properties during a 10 ns run of different proteins complexed with compound CPG	78
Table 3.3: g_sas output of the different systems.....	87
Table 3.4. A summary of interacting residues with the various ligands under study	97
Table 3.5. Drug like properties of CPs, 5PGA and ZINC hits.....	104
Table 4.1: Protein-CP complexes overall binding free energy	116
Table 4.2: The overall binding free energy (ΔG_{bind}) in $\text{kJ}\cdot\text{mol}^{-1}$ of the various proteases with 5PGA and selected ZINC compounds as determined by g_mmpbsa tool.....	117
Table 5.1: Best hits against <i>plasmodial</i> cysteine proteases identified from SANCDB.....	132

RESEARCH OUTPUTS

Publications

Musyoka, T. M., Kanzi, A. M., Lobb, K. A., and Tastan Bishop, Ö: Combined Structure- and Ligand-Based Docking and Molecular Dynamics Studies of Falcipains against a South African Natural Compound and its Analogs. *Nat. Scientific Reports*. 2016. DOI: 10.1038/srep23690.

Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Tastan Bishop, Ö: Analysis of Non-Peptidic Compounds as Potential Malarial Inhibitors against *Plasmodial* Cysteine Proteases via Integrated Virtual Screening Workflow. *J. Biomol. Struct. Dyn.* 2015. DOI:10.1080/07391102.2015.1108231

Hatherley, R., Brown, D. K., **Musyoka, T. M.**, Penkler, D. L., Faya, N., Lobb, K. A., Tastan Bishop, Ö: **SANCDB: a South African natural compound database.** *J. Cheminform.* 2015. DOI: 10.1186/s13321-015-0080-8.

Brown, D. K., Penkler, D. L., **Musyoka, T. M.** & Tastan Bishop, Ö. **JMS: An Open Source Workflow Management System and Web-Based Cluster Front-End for High Performance Computing.** *PLoS One*. 2015. DOI: 10.1371/journal.pone.0134273.

Conference proceedings

a) Oral presentations

Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Tastan Bishop, Ö: “Identification of novel inhibitors against *plasmodial* cysteine proteases using intergrated computational approaches.” *The 6th Interdisciplinary Post Graduate Conference*, Rhodes University, 8-10 October 2014.

Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Tastan Bishop, Ö: “Identification of novel inhibitors against *plasmodial* cysteine proteases using intergrated computational approaches.” *Joint SASBi-SAGS (South African Genetics and South African Society for Bioinformatics and Computational Biology) congress*, Kwalata Game Ranch, Tshwane, South Africa, 23-26 September 2014.

b) Poster presentations

Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Tastan Bishop, Ö: “Profiling the structural elements and the energetics involved in the binding of small non-peptide compounds onto *Plasmodium* and human cysteine proteases using computational approaches.” *The 23rd annual International conference on Intellelligent Systems for Molecular Biology (ISMB) and The 14th European Conference on Computational Biology (ECCB)*, Dublin, Ireland. 10-14 July 2015.

Thommas Musyoka, Aquillah Kanzi, Kevin Lobb, & Özlem Tastan Bishop: “Structural elements and the energetics involved in the binding of non-peptide compounds onto *Plasmodium* and human proteases.” *The 11^{nth} 3DSIG 2015 Structural Bioinformatics and Computational Biophysics meeting*, The Convention Centre, Dublin, Ireland. 10-11 July 2015.

Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Tastan Bishop, Ö: “In silico based methodology to determine the broad inhibitory potency of non-peptidic compounds against *plasmodial* cysteine proteases.” *Pan Africa Chemistry Network (PACN) Congress 2014 - Biodiversity and Global Challenges, A Chemical Sciences Approach*, Nov 30-Dec 2 2014, UN Conference Centre, Ethiopia.

Contributions to Publications

Musyoka *et al.*, 2016: **Combined Structure - and Ligand-Based Docking and Molecular Dynamics Studies of Falcipains against a South African Natural Compound and its Analogs.**

I performed all experiments and data analysis with an exception of the docking section. I also wrote the manuscript under the guidance of Dr. KA Lobb & Prof. Ö Tastan Bishop.

Musyoka *et al.*, 2015: **Analysis of Non-Peptidic Compounds as Potential Malarial Inhibitors against *Plasmodial* Cysteine Proteases via Integrated Virtual Screening Workflow.**

I performed all experiments and data analysis except the homology modelling and docking sections. I also wrote the manuscript under the guidance of Dr. KA Lobb & Prof. Ö Tastan Bishop.

Brown *et al.*, 2015: **JMS: An Open Source Workflow Management System and Web-Based Cluster Front-End for High Performance Computing.**

My contribution included developing the molecular dynamics pipeline and writing its section in the article.

Hatherley *et al.*, 2015: **SANCDDB: a South African natural compound database.**

Under the guidance of Rowan Hatherley and Prof. Ö Tastan Bishop together with D Penkler and N Faya I participated in the uploading and curation of ~170 compounds in the SANCDDB database.

LIST OF ABBREVIATIONS

Abbreviation	Description
3D	3-Dimensional
ACPYPE	AnteChamber PYthon Parser interface
ACTs	Artemisinin based combination therapies
AMBER	Assisted Model building with Energy Refinement
BFE	Binding free energy
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrix
CADD	Computer-aided drug design
CDKs	cyclin dependent kinases
CHARMM	Chemistry at HARvard Molecular Mechanics
CQ	Chloroquine
CQR	Chloroquine resistance
CSP	Circumsporozoite protein
DNA	Deoxyribonucleic acid
FPs	Falcipains
GROMACS	GRONingen MACHine for Chemical Simulations
GROMOS	GRONingen MOlecular Simulation
GSK	GlaxoSmithKline
HDP	Heme detoxification protein
HTS	High throughput screening
Hz	Hemozoin
IC ₅₀	Half maximal inhibitory concentration
iRBC	Infected Red Blood Cell
LBVS	Ligand based virtual screening
MAFFT	Multiple Alignment using Fast Fourier Transform
MD	Molecular dynamics
MM	Molecular mechanics
MMV	Medicines for Malaria Venture
MSA	Multiple Sequence Alignment
MSP	Merozoite Surface Protein
PBC	Periodic boundary conditions
NCBI	National Center for Biotechnology Information
PDB	Protein Data Bank
<i>PfEMP</i>	<i>Plasmodium falciparum</i> erythrocyte membrane protein
<i>PfHT</i>	<i>Plasmodium falciparum</i> -encoded facilitative hexose transporter
PROMALS3D	PROfile Multiple Alignment with predicted Local Structures and 3D constraints
PSAC	<i>Plasmodial</i> surface anion channel
PVM	Parasitophorous vacuolar membrane
QM	Quantum mechanics
R&D	Research and Development
RBC	Red Blood Cell
R _g	Radius of gyration
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation

RNA	Ribonucleic acid
ROS	Reactive Oxygen Species
SANCDB	South African Natural Compound Database
SBVS	Structure based virtual screening
SERA	serine repeat antigen
TRAP	Thrombospondin Related Anonymous Protein
WHO	World Health Organization
ZINC	Zinc Is Not Commercial

THESIS OVERVIEW

The central theme of this thesis is the use of *in silico* (computational) approaches to determine the potentiality of *plasmodial* cysteine proteases as drug targets.

The introductory chapter presents the malarial problem and the current considerations being pursued to free the world of malaria. In consideration to the continued mortality and morbidity effects from the disease, a detailed timeline of events, advances made and intrigues encountered in the fight against the disease is reviewed. As chemotherapy is still the leading approach in combatting the disease, an emphasis is given to drug resistance, which is a major hurdle in the eradication of the disease. In addition, the current state of the antimalarial drug pipeline which ensures, that efficacious and safe antimalarials are always available is reviewed. The ongoing search for compounds to eradicate malaria has recently been up-scaled through initiatives and partnerships involving private organizations, academia and pharmaceutical companies. In this chapter computational technology, genomics and structural biology in modern drug development and discovery are also reviewed. Genomics initiatives and bioinformatics in the discovery and elucidation of new drug targets are also reviewed. A major advance is the identification of potential drug targets that can be considered in the development of new antimalarial drugs.

This thesis mainly focusses of falcipains (FPs) and homologs, cysteine proteases used to degrade haemoglobin and other proteins to their constituent amino acids in *Plasmodium* species. FPs have been identified: FP-1, FP-2, FP2' and FP-3. FP-2 and FP-3 are considered valid drug targets.

Chapter 2 mainly focuses on identifying homologs (related proteins) of FP-2 and FP-3 from other *Plasmodium* species that infect humans and laboratory model organisms such as mice or rats. As these proteins are closely related to human cathepsins, BLASTP search tool was

used to identify human homologs. Understanding protein structure is key in elucidating its mechanism of action, and how it can be therapeutically targeted. Thus, several *in silico* approaches were used to determine differences in the sequence, structure, evolution and physicochemical properties between the two classes of proteases (*plasmodial* proteases and human cathepsins). Several sequence, structural and physicochemical similarities and differences were established between the *plasmodial* and human proteins and also within the individual classes.

Chapter 3 describes the use of GROMACS, a molecular dynamics (MD) simulation software to study the time depended stability and evolution of protein-ligand complexes. This was a continuation of previous work by a former MSc student who performed docking studies using two sets of compounds: Cyanopyrimidine nitrile derivatives (CPs) and a South African natural compound (5PGA) and its Zinc Is Not Commercial (ZINC) analogs. For the CPs, 10 nanoseconds (ns) simulations were performed while for the 5PGA and its analogs the simulation runs were for 20 ns. The dynamic ligand binding process was analysed. This led to the identification of residues critical to ligand binding of the ligands and the binding modes of the different ligands. Additional work describing a pipeline to allow for the automation of MD simulation previously published is also described.

Chapter 4 mainly describes the quantification of the energetics between the different protein-ligand systems studied in Chapter 3 using MD simulations. Using single trajectory approach method through the `g_mmpbsa` tool, the overall binding free energy (BFE) between each protein-ligand complex is studied. The overall BFE was decomposed to determine the individual contributions of the van der Waals, electrostatic, polar solvation and solvent accessible surface area energy terms. In addition, the individual contribution of each amino acid (aa) in the proteins was determined to identify key binding residues. Part of this work

and corresponding results from Chapter 3 has been published in the Journal of Biomolecular Structure and Dynamics while the rest in the Nature Scientific Reports Journal.

Chapter 5 is work in progress including the mining of non-peptide natural compounds with inhibitory activity against *plasmodial* proteases from the South African Natural Compounds Database (SANCDDB). This is a fully referenced growing chemical database developed by members (myself included) of Research Unit of Bioinformatics (RUBi) and has previously been published. Presently, potential hits against FP-2 and FP-3 and their homologs from *P. knowlesi*, Knowlesipain-2 (KP-2) and Knowlesipain-3 (KP-3) have been identified. The newly identified hits are predicted to have better inhibition potencies compared to previously tested compound from South Africa against the already tested proteins.

Finally, Chapter 6 summarizes up the major findings while also presenting the future prospects.

CHAPTER 1

Malaria

Malaria is a two word name with a Latin prefix 'mala' and Italian suffix 'aria' together meaning bad air. It was initially referred to as the "marsh fever" as it was associated with marshland or swampy areas. Malaria is an infectious disease caused by a group of obligate single-cell parasitic organisms of the Plasmodium genus and remains a serious global health problem. In 2014, the World Health Organization (WHO) estimated that about 97 countries and territories had ongoing active transmission of malaria with 3 billion people being at a risk of infection globally. The majority of these areas are in tropics of Africa, Asia and South America. In 2013, ~0.5 million people died of malaria with Africa accounting for more than 90% of these. This chapter reviews malaria, the global strategies adopted to eliminate the disease, the major hurdles in achieving a malaria free world and the current state of the antimalarial drug discovery pipeline. Technological advances in research and development (R&D) of drugs and the resulting gains will also be discussed.

1.1 Malaria: The ever elusive global health challenge

Malaria remains an exigent problem in global public health with roughly half of the world population living in malaria endemic regions, mainly tropical and subtropical regions of Africa, South America and Asia¹. Malaria is the most prevalent and severe tropical disease causing more than a half a million deaths annually, 90% of which are pregnant mothers and children under five years from Africa¹. Malaria is caused by a group of obligate erythrocytic protozoan parasites of the genus *Plasmodium* that infect vertebrate hosts such as reptiles, birds and mammals. Each *plasmodial* species however, has a narrow host range such that *in vivo* infection studies of species targeting humans can only be performed on primates². Transmission occurs after a bite from an infected female mosquito from ~30 *Anopheles* species, which introduces sporozoites into the host blood stream. In humans, malaria is caused by five distantly related species *viz.* *P. falciparum* (*Pf*), *P. vivax* (*Pv*), *P. ovale* (*Po*), *P. knowlesi* (*Pk*) and *P. malariae* (*Pm*)^{1,3,4}. *Pk* was originally known as the malarial pathogen of the pig-tailed (*Macaca nemestrina*) and long tailed macaques (*Macaca fascicularis*) but has recently been reported to infect humans in Asia^{5,6}. The initial forms of *Pk* have close morphological similarity with *Pm*⁷. However, it has a shorter life cycle of 24 hours (hrs) compared to *Pm* such that delayed treatment may lead to life threatening complications due to an increased parasite load in the host blood stream and kidney dysfunction⁸. Highly adaptable *Pf* is the most virulent species accounting for most death and disease cases in Africa^{1,3,4}. It binds to the endothelium lining during the erythrocytic phase of infection and accumulates in organs including the brain⁹. Evolutionarily, *Pf* and its closest relative *P. reichenowi* (*Pr*), which infects chimpanzees, form a separate taxonomic group known as the hominid clade². *Pv* is the mostly widely distributed human parasite. However, the prevalence of the Duffy negative trait in African populations lowers the threat of *Pv* which causes benign malaria in temperate regions of the world. *Po* and *Pm* are less prevalent and less lethal than *Pf* and *Pv*.

In addition to the human-specific *plasmodia*, several species that infect non-human laboratory animals used as models to understand the parasite biology, host-parasite interactions and for antimalarial drug development¹⁰. These include *P. berghei* (*Pb*), *P. chabaudi* (*Pc*) and *P. yoelii* (*Py*) that infect mice and rats.

1.2 *Plasmodium* life cycle

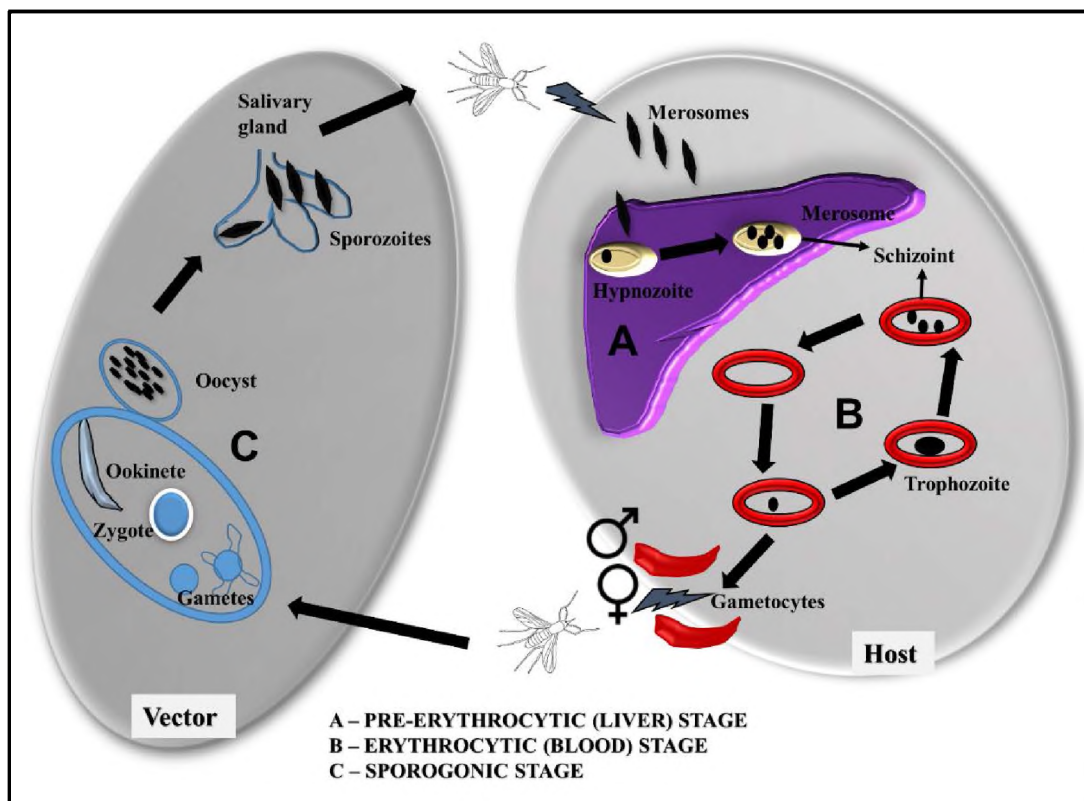


Figure 1.1: *Plasmodium* parasite life cycle. A schematic representation of the *plasmodial* life stages in vector and host. During mosquito feeding, sporozoites are inoculated into the host circulatory system by an infected female *Anopheles* mosquito and invade hepatocytes (A). Sporozoites then divide into haploid merozoites which are released back into the circulatory system to initiate the blood stage (B). Merozoites invade RBCs where they reproduce asexually leading to the release of thousands of merozoite progeny which invade uninfected erythrocytes. A fraction of the circulating merozoites in iRBCs develop to sexual gametocytes for ingestion by a feeding vector. In the midgut of the mosquito, the gametocytes mature into gametes marking the beginning of the sexual or sporogonic phase (C). After fertilization, a diploid zygote is formed which develops to an oocyst through a mobile ookinete. The oocysts grow and divide into thousands of active haploid sporozoites. Finally the oocyst ruptures releasing sporozoites into the vector hemocoel from where they migrate into the salivary glands for transmission to a vertebrate host. Adapted from Winzeler EA, 2008¹¹.

Plasmodium is a multi-stage protozoan with an intricate life cycle alternating between an arthropod vector and a vertebrate host such as humans^{12,13}. The incubation period varies between seven and 30 days depending on the *plasmodial* species with *Pf* having the shortest and *Pm* the longest. The life cycle consists of three phases: a pre-erythrocytic (liver or hepatic), an erythrocytic (blood), and a sporogonic (vector) stage accompanied by a series of morphological and biochemical transformations (Figure 1.1). During the liver and the blood stage the parasite life is intracellular in hepatocytes and erythrocytes respectively. The sporogonic or asexual reproductive stage occurs in the infected female mosquito giving rise to *plasmodial* sporozoites which are injected into the host blood stream during a blood feeding sessions.

To avoid blood clotting and pain, the mosquito saliva contains anti-hemostatic enzymes and anti-inflammatory chemicals¹⁴. Sporozoites are nucleated highly motile cells with a single mitochondrion, an apicoplast and a single microtubule interconnected by tethering proteins. In the blood stream, sporozoites migrate to the liver, cross several Kupffer cells and invade the hepatocytes¹⁵. Sporozoites carry “sporozoite surface proteins” such as thrombospondin-related anonymous protein (TRAP) and circumsporozoite protein (CSP) thought to function in recognition and anchoring during the hepatocytes entry process¹⁵. In the hepatocytes, sporozoites proliferate asexually and form merozoites containing thousands of haploid spindle-shaped merozoites. Depending on the parasite species, this process may take weeks to months. In *Pf* and *Pm*, the maturation process takes one to two weeks. In *Pv* and *Po* sporozoites remain in the human hepatocyte for months to years as hypnozoites before maturing and causing late malaria lapses¹⁶. The merozoites are released into the blood stream where they disintegrate releasing merozoites, which in turn invade erythrocytes initiating the erythrocytic stage. During the pre-erythrocytic stage, the disease has no clinical

manifestations of the disease. The host immune defence mechanisms are thus not activated allowing the circulating merozoites to survive.

The erythrocyte invasion process by merozoites is mediated by an association of proteins from both the merozoites and host red blood cell (RBC). Proteins involved include the merozoites surface proteins (MSP) 1, 7 and 9 which in turn bind to the erythrocyte band 3 protein^{17,18}. At the apex of the merozoites are rhoptries and micronemes, a mixture of proteinases and metabolic enzymes that help in the invasion process through the erythrocyte membrane. In *Pv*, invasion requires the recognition of the Duffy blood group antigen, a known receptor for the Interleukin-8 (IL-8). After invasion, the erythrocytic stage begins, and merozoites undergo a trophic period during which they enlarge losing their apical rings, conoid, and rhoptries structures while their nuclei become lobulated. The early trophozoite is referred to as 'ring form'. Its enlargement is accompanied by active metabolism involving ingestion of host cytosol and degradation of haemoglobin into its constituent aa residues. A by-product of the degradation is iron containing protoporphyrin IX which could generate potentially toxic reactive oxygen species (ROS) through the Fenton reaction^{19,20}. Despite lacking heme oxygenases, *Plasmodium* parasites convert the free heme to unreactive dark, crystalline hemozoin (Hz). Lin *et al.*, (2015) established that *Pb* mutant parasites lacking haemoglobin degradation enzymes still develop into schizonts and gametocytes without forming Hz²¹. The trophic stage is characterized by multiple rounds of nuclear division without cytokinesis leading to round schizonts, each of which contains 12-16 merozoites⁹. Upon rupture of the infected red blood cell (iRBC), the merozoites are released into the blood stream to initiate another round of replication.

The invasion of RBC, degradation of haemoglobin and the ultimate rupturing are dependent on proteases²². These proteases constitute the main focus of this thesis. To avoid parasite clearance by the spleen, *Pf* has evolved a mechanism known as cytoadherence whereby

trophozoites and schizonts bind host endothelial receptors such as *Pf* erythrocyte membrane protein (*Pf*EMP) remaining in capillary venules leading to malaria complications such as cerebral and placental malaria²³. Rupturing of iRBC releases Hz. Hz, initially thought to be a metabolic waste product only, can accumulate in the lung, liver, brain and spleen leading to malaria immunopathogenesis²⁴. The erythrocytic stage is responsible for the clinically observed pathological symptoms such as cycles of fever paroxysms, nausea, abdominal and back pains. *Pf* infection can also lead to acute renal failure, cerebral malaria, metabolic acidosis, hemoglobinuria, blood coagulation abnormalities, hypoglycemia and acute respiratory distress syndrome (ARDS)²⁵. A small proportion of released merozoites develop into male (microgametocytes) and female (macrogametocytes) gametocytes, the sexual forms of the parasite, although the mediators of this process largely remain unknown. The sexual cycle or gametogenesis starts with gametocyte ingestion by a mosquito. Male gametocytes form microgametes through exflagellation while the female gametocytes form the macrogametes. In the mosquito mid gut, the microgametes fertilize the forming a zygote, which later develops into a motile ookinete which penetrates the gut epithelial cells and forms an oocyst. The oocyst undergoes several rounds of asexual replication producing sporozoites released into the mosquito hemocoel and from where they eventually migrate to the mosquito salivary glands to inoculate a new human host during the next blood meal perpetuating the parasitic life cycle (Figure 1.1).

1.3 Antimalarial drug history

1.3.1 The role of natural products in antimalarial drug discovery

Malaria has killed humans throughout recorded history. Many advances have been made in developing drugs against malaria. Natural products provided lead compounds for different drugs²⁶⁻²⁹. Natural products have complex molecular architectures with unique arrangement

of functional groups. Their structural diversity and chemical properties are critical to drug discovery. The main antimalarial drugs have all been developed from natural products³⁰.

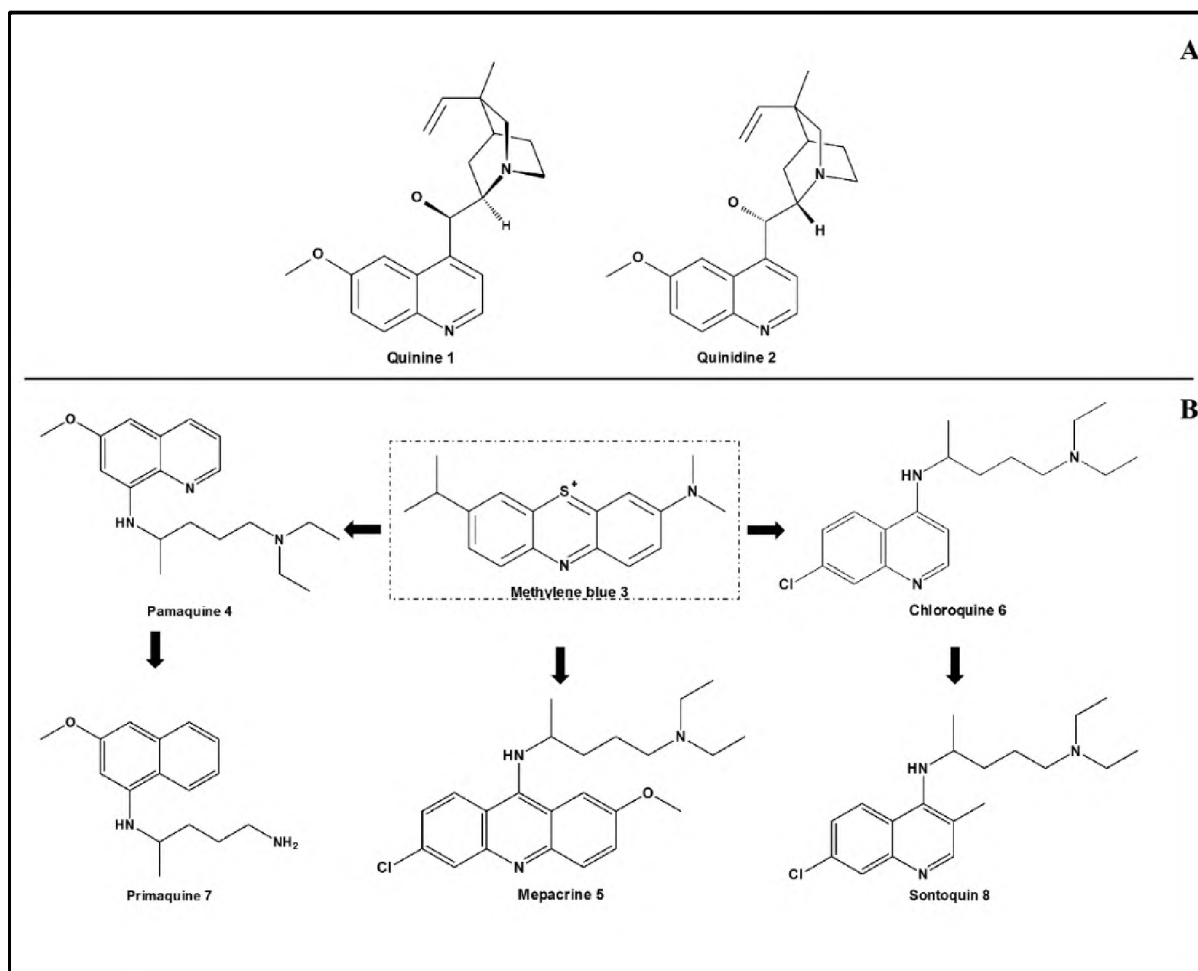


Figure 1.2: 2D chemical structures of the early antimalarial drugs. A) *Plasmodicidal* natural 4-methanolquinoline alkaloids quinine (1) and quinidine (2) from the *Cinchona* tree. B) The first synthetic forms of antimalarial drugs from Methylene blue (3). Adapted from Schlitzer M, 2007³¹. Structures were drawn using ChemDraw Ultra 10.0³².

In the 1620s, Jesuit missionaries discovered the healing powers of cinchona tree (*Cinchona calisaya*) bark in Peru and Bolivia forests. A century later, the active compounds quinine (1) and cinchonine (2) were isolated by Caventou Joseph and Pierre Pelletier (Figure 1.2)³³. Due to war related embargos and the heavy demand for cinchona, William Henry Perkins attempted to synthesise quinine in 1856. He synthesized textile dye named ‘mauve’ starting a dye industry in Germany, several of which were used in medicine while microbiologists used

them to identify and classify organisms. In 1880, *Plasmodium* itself was discovered by Charles Louis Alphonse, a French army surgeon. In 1886, Camillo Golgi established that there were two forms of malaria based on the symptoms and number of merozoites in the blood stream. The first species to be classified were *Pv* and *Pm* in 1890 by Giovanni Batista and Raimondo Filetti³⁴. Mosquitoes were identified as the vector fifteen years later by Ronald Ross³⁵. In 1891, Paul Ehrlich used methylene blue (**3**) to stain malaria parasites and proposed that the dye could be used to kill the parasites. His hypothesis was confirmed when he cured two malarial patients using methylene blue³⁶.

In the 1920s, scientists at the chemical company Bayer modified methylene blue by replacing its side methyl groups with unique heterocyclic groups. This created the first aminoquinoline named plasmoquine or pamaquine (**4**) and several other derivatives³⁷ (Figure 1.2). Due to its severe toxic side effects, pamaquine was discontinued in 1925 and replaced by primacune (**7**) three decades later, a more tolerated derivative. Another derivative mepacrine (**5**) or trade name Atebrin® was obtained by fusing the diethylaminoisopentylamino side group with an acridine heterocycle. At the time, Japan had invaded Indonesia blocking the supply of quinine. American, Australian and British scientists collaborated to develop novel medicines to fight malaria, a threat to their soldiers. About 20,000 quinine derivatives were synthesised and tested identifying a far superior compound named as resoquin. Although initially abandoned due to severe side effects, it was later reviewed and finally accepted in 1946 under new name, chloroquine [CQ] (**6**) eventually replacing quinine³⁸. CQ is highly effective in clearing the asexual blood forms of major parasites, but it was found to have a very small therapeutic index with a 30 mg/kg being considered lethal³⁹.

For more than two decades after its discovery in 1946, chloroquine was the most successful antimalarial drug and became a mainstay for malaria treatment. The drug was initially mixed

with table salt and distributed an approach commonly referred to as the “Pinotti’s method”⁴⁰. A decade after CQ was introduced, resistant malarial (CQR) strains were detected in the Cambodia and Thailand. The wide-spread application of sub-therapeutic doses of CQ are believed to have led to the emergence of drug resistance⁴¹. Today CQR strains are present in essentially all regions with active malaria.

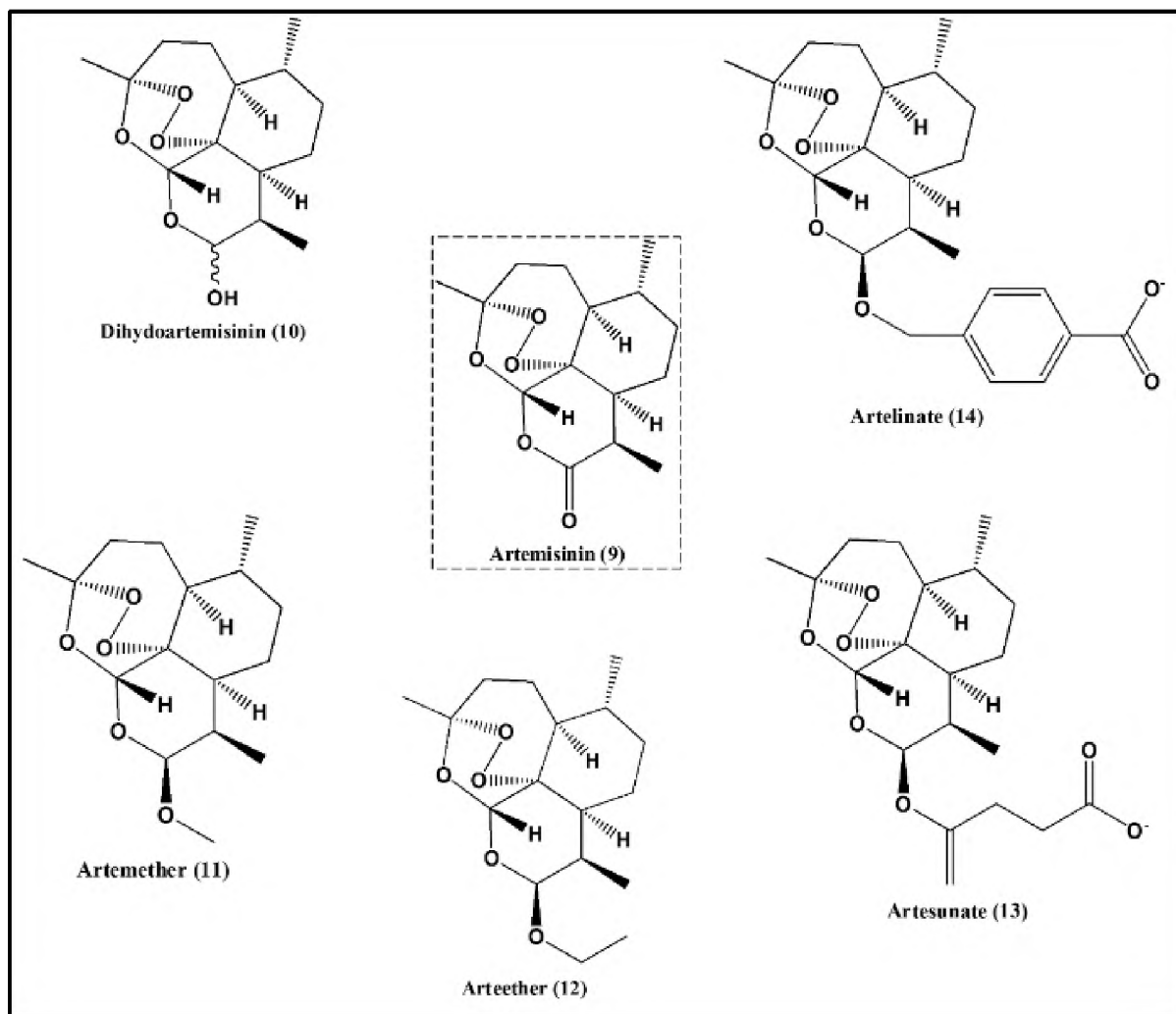


Figure 1.3: Artemisinin and its derivatives which are the basis of current antimalarial therapies. Boxed is Artemisinin, the parent compound of all artemisinin based combination therapies (ACT) derivatives.

The search for new drugs to address CQR identified an interesting new natural product from *Artemisia annua* (sweet wormwood or qinghao) leaves by Chinese scientists in 1971 led by

the 2015 Nobel laureate of medicine and physiology Youyou Tu. This was named “qinghaosu” translating to the essence of qinghao⁴². In 1972, the pure crystalline form of the sesquiterpene lactone was isolated and its structure reported in 1979 under the name artemisinin (**9**)^{43,44}. Due to its higher therapeutic index than CQ and its efficacy in clearing CQR strains, artemisinin was considered a “magic bullet” against malaria^{45,46}. Due to its inherent poor bioavailability, semisynthetic derivatives *viz.* dihydroartemisinin (**10**), artemether (**11**), arteether (**12**), artesunate (**13**) and artelinate (**14**) have been developed (Figure 1.3)⁴⁷.

The unique endoperoxide group or the 1,2,3-trioxane scaffold is the basis for the antimalarial activity of artemisinin and its derivatives. Under resolution WHA60.18, the WHO in 2007 banned the use of artemisinin-based monotherapies, and recommended the use of artemisinin based combination therapies (ACTs) to maintain the efficacy of artemisinin⁴⁸. Since 2015, five ACTs have been approved by the WHO for the use to treat uncomplicated malaria cases^{49,50}.

1.3.2 Other antimalarial natural products

Owing to the importance of natural products as source of leads for the development of drugs against diseases, a well-documented literature of plants with antimalarial activity exists^{26,27,30,51}. Many communities in the tropics have identified plants whose crude extracts are used to treat malaria. However, the safety and efficacy of such extracts need to be checked in order to determine their potential as source of lead compounds for antimalarial drugs.

Using high-throughput screening systems (HTS) and *in silico* approaches, many research groups are determining the active compounds of various plant extracts. This has given rise to several private and public relational databases of natural compounds such as NAPRALERT⁵²,

TCM-ID⁵³, NuBBE_{DB}⁵⁴, ConMedNP⁵⁵, AfroDB⁵⁶ and the South African Natural Compound Database (SANCDDB)²⁹.

A 2003 review by Saxena *et al.*, listed 127 alkaloids, 18 quassoids, 27 triterpenoids, 23 sesquiterpenes, 21 flavonoids and xanthenes, 9 quinones and 25 assorted compounds with potential antimalarial activity⁵⁷. Another listed 31 indole alkaloids with *in vitro* and *in vivo* anti *plasmodial* IC₅₀ values in the μM range and desirable selectivity⁵⁸. Another review focusing on the ten year window from 1998-2008 of natural compounds provided 266 anti *plasmodial* compounds from a variety of chemical classes⁵⁹. Several natural compounds with high *in vivo plasmodia* clearance have been identified and may constitute novel antimalarial drugs once approved. These include borrelidin, a natural antibiotic compound that inhibits threonyl-tRNA synthetase⁶⁰ and a trioxolane with a peroxide pharmacophore which has also shown promising results in clinical trials. It targets all asexual blood stages of *Pf*⁶¹.

1.3.3 Synthetic antimalarial drugs

Antimalarials from natural sources such as cinchona and sweet wormwood trees raise concerns as supply of their bioactive components cannot be sustained overtime. Various synthetic drugs based on quinine or artemisinin have been developed. The continued use of a single drug leads to *plasmodial* drug resistance. The adoption of synthetic chemistry to modify the main antimalarial prototypes and synthesize novel derivatives is critical. This is to address the drug resistance, increase the bioavailability, half-life, efficacy and safety of antimalarials.

1.3.3.1 Classification of available antimalarial drugs

Malarial drugs are classified either to reflect the stage in the malarial life cycle they target or their chemical structure and function (Table 1.1).

Table 1.1: Classification of known antimalarial drugs^{62,63}

Type	Class	Details
<i>Plasmodial cycle target</i>	Primary tissue schizontocides	Target pre-erythrocytic forms of <i>Pf</i> and <i>P</i> . Not widely used since it is hard to predict malarial infection before the erythrocytic symptoms. Include pyrimethamine and primaquine.
	Blood schizontocides	Target the blood asexual forms of all malarial species. Most common antimalarials. Include mefloquine, quinine, mecaprine and 4-aminoquinolines, sulfones, tetracycline, halofantrine).
	Gametocytocides	Target the sexual forms of the malarial parasite to block infection to the vector. Include the 8-aminoquinolines.
	Sporontocidal	Block the sporogonic phase of the parasite cycle thus preventing infection to the host. Include proguanil, pyrimethamine and chloroguanide.
	Secondary tissue schizontocides	Target the hypnozoites of <i>Po</i> and <i>Pv</i> preventing a relapse of reactivation of residual parasites. Administered as a follow up of the treatment after a primary attack. Include pamaquine, and primaquine).
<i>Chemical structure and or function</i>	8-aminoquinolines	First synthetic antimalarial primaquine and its derivatives tafenoquine and bulaquine.
	4-aminoquinolines	Developed to overcome chloroquine resistance. Include amodiaquine, ferroquine, isoquine, pyronaridine and naphthoquine
	Aryl amino alcohols	The natural alkaloid quinine and derivatives mefloquine, halofantrine and lumefantrine.
	Folate synthesis inhibitors	Type 1: Competitive dihydropteroate synthase inhibitors. Include sulfachrysoidine and sulfadoxine. Type 2: Inhibit dihydrofolate reductase. Include pyrimethamine (fansidar), cycloguanil and dapsone.
	Protein synthesis inhibitors	Target the protein synthesis in apicoplasts. Include tetracycline, doxycycline, clindamycin, azithromycin, macrolides lincosamides, chloramphenicol and thiazole antibiotics.
	Peroxides	Artemisinin and its derivatives. Mode of action not clear but thought to involve reactive C-radicals. Include artemether, arteether, artesunate, artelinic acid and dihydroartemisinin.
	Respiratory chain blockers	Mostly naphthaquinones. Include atovaquone and buparvaquone.
	Iron chelators	Deplete iron for metabolism Include desferrioxamine.

1.4 Antimalarial drug resistance

As a result of co-evolution within their hosts, *plasmodia* have acquired complicate strategies and molecular structures similar to those of the host. Consequently, the selective targeting of the parasites has been an extremely daunting task and thus continue to exercise their witty nature as colonists to their hosts^{64,65}. Global elimination of malaria is greatly hampered by the parasite ability to develop resistance to all available drugs⁶⁶. Major drug resistance involves single or multiple aa substitutions due to insufficient drug levels. Mutations decrease affinity to drugs, amplify target production, decrease drug activation or change drug accessibility.

South East Asia has seen major drug resistance develop due to low immunity and non-compliant use of drugs⁶⁷. The analyses of *Plasmodium* resistant strains by molecular, genetic and pharmacological techniques have identified key mutations in enzymes or transporters that confer drug resistance (Table 1.2).

Table 1.2: Key genes in *P. falciparum* and aa mutations leading to drug resistance

Drug	Gene (Mutation)	Reference
Chloroquine	<i>Pfcr</i> (K76T), <i>Pfcg2</i> , <i>Pfcm</i> <i>dr1</i> (N86Y)	68
Quinine	<i>Pfcr</i> (K76T), <i>pfmdr1</i> (N1042D), <i>pf</i> <i>nh</i> <i>e1</i>	69–71
Atovaquone	Cytochrome <i>b</i> gene (Y268S)	72
Amodiaquine	<i>Pfcr</i> (K76T), <i>Pfcm</i> <i>dr1</i> (N86Y)	73
Mefloquine	<i>Pfmdr1</i>	73
Sulfonamides	<i>Pfdhfr</i> , <i>pf</i> <i>dhps</i>	74
Lumefantrine	<i>Pfcr</i> , <i>pfmdr1</i>	75
Piperaquine	<i>Pfcr</i>	76
Antifolates	<i>Pfdhfr</i> (N51I , C59R , S108N , and I164L)	77
Artemesinins	<i>Pfkelch13</i> (C580Y), <i>PfATPase6</i> (A623E)	78

Pfcr = *Pf* chloroquine transporter, *Pfmdr1* = *Pf* multidrug resistance transporter1, *Pfdhfr* = *Pf* dihydrofolate reductase, *Pfdhps* = *Pf* dihydropteroate synthase and *Pfnhe1* = *Pf* sodium/proton exchanger 1

Pb mutants lacking haemoglobin degrading enzymes mature to schizonts and gametocytes²¹ and are resistant to CQ with major implications for drug development especially against *Pv* and *Po* that grow inside reticulocytes.

1.5 Antimalarial drug development in the post genomic era

Genome sequencing technologies over the past two decades have determined genome of humans (2001)⁷⁹, *Pf* (2002)⁸⁰ and *Anopheles gambiae* (*Ag*) in 2002⁸¹, a major boost to the field of medical sciences^{82–84}. Before, drug development was laborious as target identification required cellular studies, pharmacological models and protein biochemical assays. The availability of appropriate genomes has allowed comparative genomics and data mining identify druggable targets central to the antimalarial drug discovery process.

In 1991, Nobel laureate Walter Gilbert predicted of paradigm shift in biological research in the 21st century as genes would be available in electronic databases⁸⁵. Accordingly genomes

of many important organisms have been assembled and annotated. Hybrid computational methods and conventional drug discovery techniques are being used to develop antimalarial drugs. Although a range of drugs have already been developed in the past, more are required to overcome malaria. Genomic and protein structural data allow drug targets unique to the parasites to be focussed on^{86,87}.

1.5.1 Establishing a drug target in antimalarial drug development

A binding pocket in a molecular structure does not automatically guarantee the suitability of this structure as a drug target. In this thesis, a drug target will be defined as an essential protein (receptor, channel, transporter or enzyme) involved in the signalling, transport or metabolic pathway whose inhibition may lead to the elimination of the parasite. It is thus considered to be essential in the survival of the parasite. To validate a potential drug target, a biological explanation of its therapeutic value is critical. In addition, druggability profile of the target must be defined to indicate that drug-like molecules can target the binding site with sufficient affinity and specificity. An ideal binding pocket should be buried to increase interaction surface and be of appropriate size to accommodate a drug-like compound. Several methods for assessing druggability have been established that rely on sequence, structure, and precedence. Potential drug targets are mostly proteins of essential metabolic pathways or transport channels. Bioinformatics together with systems biology, microarray studies, comparative proteomics, transcriptome studies, gene network studies, structural studies, molecular docking, molecular dynamics and evolutionary studies may be used to identify new drug targets⁸⁸⁻⁹¹.

1.5.2 Antimalarial drug targets

Differences between metabolic pathways of host and *Plasmodium* offers a plethora of drug targets that can be utilised for novel antimalarial drug development^{92,93}. Of great interest is events following *plasmodia* invasion of erythrocytes which include heme degradation and

detoxification, fatty acid biosynthesis, nucleic acid metabolism, and the oxidative stress⁹³. Similarly, the blood stage has attracted significant interest. After invading an erythrocyte, *Plasmodium* degrades up to 75% of haemoglobin to its constituent aa. It does so as 1) a source of essential aa and energy as the parasite lacks the ability to synthesize them 2) to regulate osmotic pressure and 3) to create space for growth and replication^{94,95}. Many endo- and exo-peptidases participate to degrade α and β chains of haemoglobin including proteases (aspartic, cysteine, serine, threonine, metallo and mixed), plasmepsin I, II, III and IV and aminopeptidases^{96,97}. Due to their critical importance, this thesis focuses on *Pf* falcipains (FPs), homologs from other *plasmodial* species as well as human homologs. These enzymes belong to the group of cysteine proteases, and will be discussed at length in the following chapters. Also of a major attention are the enzymes involved in converting haemoglobin to hemozoin. These include histidine-rich proteins, heme detoxification protein (HDP), heme binding proteins and reduced glutathione molecule^{24,98}. The parasite primes the RBC membranes by introducing specialised ion and transport channels for fuel uptake and waste disposal. Only hours after post- infection, iRBC will contain both parasitophorous vacuolar membrane (PVM) and *plasmodial* surface anion channel (PSAC) which are commissioned for nutrient acquisition^{99,100}. Several transporters for nutrient uptake including *Pf*-encoded facilitative hexose transporter (*Pf*HT) are localised in the parasite's plasma membrane to take up glucose its primary source of energy during the intra-erythrocytic stage of the parasite development.¹⁰¹.

A second major by-product of haemoglobin degradation alongside heme that could kill the parasite^{20,24,102}. For a redox equilibrium, *plasmodia* utilise a redox system involving several enzymes *viz.* glutathione synthetase, glutathione reductase, glutathione-S-transferase, superoxide dismutase, γ -glutamyl-cysteine synthetase, glutamate dehydrogenase and

thioredoxin reductase¹⁰³⁻¹⁰⁷. These enzymes are correspondingly important drug targets and endoperoxide antimalarial drugs such as artemisinin works by inducing oxidative stress⁴⁷.

As *plasmodia* are auxotrophic for purine bases, they must source host nucleotides for DNA and RNA during growth and replication as the parasites lack a *de novo* biosynthetic pathway of purines (nucleotide building blocks)^{108,109}. On the contrary, *plasmodia* must synthesise pyrimidine nucleotides from the scratch as they cannot salvage host's pyrimidine bases^{108,109}. Hence, the enzymes; inosine dehydrogenase, hypoxanthine-guanine phosphoribosyl transferase and adenylosuccinate synthase^{110,111}, carbamoyl phosphate synthase, dihydroorotase, aspartate transcarbamylase, orotatephosphoribosyl transferase, dihydroorotate dehydrogenase, and orotidine 5-phosphate decarboxylase^{112,113} involved in the purine salvage and pyrimidine biosynthetic pathways respectively are important drug targets.

Another requirement of the growing parasites is phospholipid biosynthesis for membranes¹¹⁴. The parasites utilise two pathways: *de novo* choline pathway also known as Kennedy pathway where choline kinase is a major antimalarial drug target and the serine decarboxylation-phospho-ethanolamine methylation pathway¹¹⁵.

The asexual division and replication stages of malaria parasites are mediated by a group of enzymes known as cyclin dependent kinases (CDKs). Comparing sequences of *Plasmodium* and host CDKs reveal aa residue substitutions especially at their binding pocket which can be potential drug targets¹¹⁶.

The invasion and subsequent release of the parasite yield further potential drug targets. To invade RBCs, *plasmodia* merozoites utilise an array of invasion apparatuses first to degrade the protective cage known as parasitophorous vacuole and subsequently the erythrocyte cell membrane^{17,92,117}. Several secondary interactions such as the Duffy binding like proteins that mediate the invasion process have been identified, and are attractive targets for the

development of protective vaccines. In *Pf*, signal peptide peptidase (SPP) which is an aspartyl protease binds to RBC band 3 receptor marking the onset of the invasion process¹¹⁸. Through chemical studies, inhibiting SPP using L-685 leads to inhibition of the invasion process and the resulting downstream processes including replication and release of parasite progeny¹¹⁸. After replication, the iRBCs rupture releases mature merozoites. Several proteases have been identified as key players in degrading the iRBCs membrane and leading to its eventual rupturing. These include the serine repeat antigen (SERA), cysteine proteases like FPs and dipeptidyl peptidase 3 (DPAP3)^{96,119}. Several cysteine protease inhibitors such as calpeptin, leupeptin (a general serine and cysteine protease inhibitor) and E-64 ((1S,2S)-2-(((S)-1-((4-guanidinobutyl)amino)-4-methyl-1-oxopentan-2-yl)carbamoyl)cyclopropanecarboxylic acid) have been identified. However, the named inhibitors are yet to be approved as drugs^{120,121}. Hence, a complete understanding of the complex biochemical pathways should offer novel solutions in the design and discovery of novel drugs.

1.5.3 Drug development pipeline and the place of modern computer technology

Drug R&D process is an extremely protracted and expensive enterprise with the development of a single drug taking up to 10-15 years and US\$ 500-800 million¹²²⁻¹²⁴. This forces pharmaceutical companies to concentrate on drugs with guaranteed fiscal returns. In 2011, only ~3% of pharmaceutical research budgets were targeted to controlling major infections¹²⁵⁻¹²⁷. Besides the expensive nature, the process is aggravated by the numerous failures. For every 7,500 drug candidates entering the development pipeline, only one or none gets to the approval. Over the last three decades, the R&D process has evolved tremendously from basic science characterised by trial and error to a more complicated interdisciplinary approach that is rational. Computer-aided drug design (CADD) was firstly covered by Fortune magazine (October 5th 1981) under the title “Next Industrial Revolution: Designing Drugs by Computer at Merck”¹²⁸. Since then, more technological innovations embedded on

several disciplines *viz.* biology, chemistry, pharmacology, mathematics, computer science, and molecular modelling have been developed with substantial gains being witnessed in the field of pharmaceuticals^{129–131}.

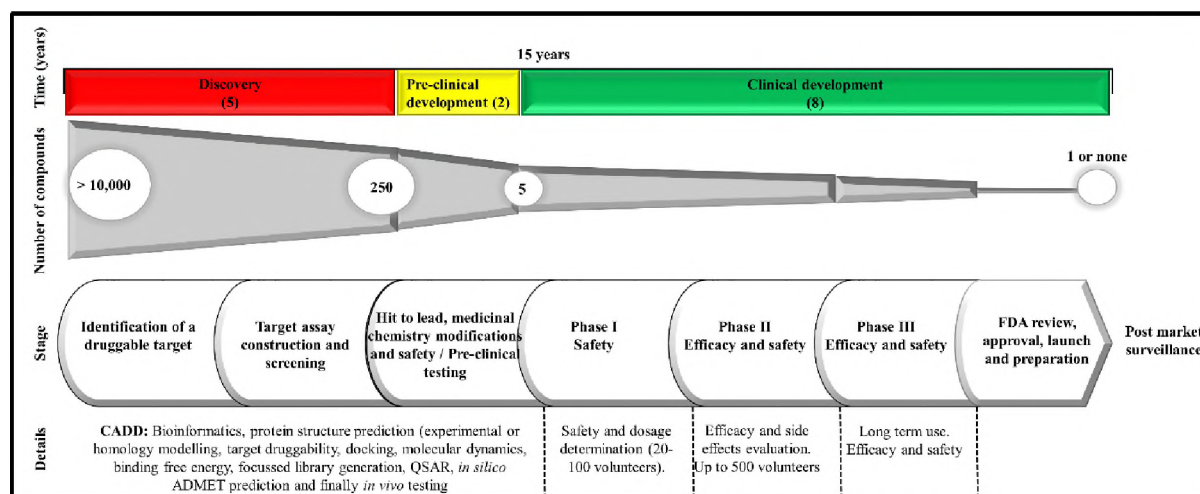


Figure 1.4: Modern drug discovery pipeline. A representation of the sequential steps, cost estimate, timeline of events and major activities involved in the drug discovery and development. Modified from Roses AD, 2008¹³²

1.5.4 The current status of the antimalarial drug pipeline

BIO Ventures for Global Health¹³³ in 2014 estimated that, 43 drugs were being developed to fight malaria infections. Majority of these drugs are targeted only against *Pf.* Up to six formulations, mostly based on artemisinin, were undergoing phase III clinical trials and eight phase II. Table 1.2 lists the names of drug candidates in pre-clinical and clinical trials. The number of compounds entering the pipeline have increased over the past years mostly due to public-private partnerships (PPP's)¹³⁴ involving pharmaceutical companies (GSK and Novartis), universities and non-profit organizations. Medicines for Malaria Venture (MMV) started in 1999 is a prime example. HTS initiatives have identified 20,000 compounds with antimalarial activity. They are deposited in the ChEMBL neglected tropical diseases archive¹³⁵. MMV has also established Malaria Box – a representative selection of 400

compounds obtained from the >20,000 hits to further promote the investigation of these compounds¹³⁶.

Table 1.3: Antimalarial drug candidates undergoing (pre-) clinical trials. Bold are drug candidates with new indications and regimens (adapted from BIO Ventures for Global Health)¹³³

Phase	Compound
Pre-clinical	BCX4945 (<i>Pf</i>), CDRI 99/411 (<i>Pf</i>), LMK235 (<i>Pf</i>), JPC-2997(<i>Pf</i>), MMV121 (<i>Pf</i> , <i>Pv</i>), NPC1161B (<i>Pf</i> , <i>Pv</i>), P218 DHFR inhibitor (<i>Pf</i>), Pyrazoles 21A092 (<i>Pf</i>), Quinolones (<i>Pf</i> , <i>Pv</i>), Reversed chloroquine molecules (<i>Pf</i>), Trioxaquinones (<i>Pf</i>), Trimethoprim/Sulfamethoxazole (<i>Pf</i>), OZ439-Ferroquine (<i>Pf</i>), OZ439-Piperaquine (<i>Pf</i>), Tinidazole (<i>Pv</i>)
Phase I	ACT451840 (<i>Pf</i>), Aminopyrdinel (<i>Pf</i>), Allocryptopine-Protopine-Berberine (<i>Pf</i>), CDRI/63 + CDRI 97/98 (<i>Pf</i>), GSK369796 (<i>Pf</i>), SJ733 (<i>Pf</i>)
Phase II	AQ-13 (<i>Pf</i>), Artemisone (<i>Pf</i> , <i>Pv</i>), DSM265 (<i>Pf</i>), Ferroquine (<i>Pf</i>), Fosmidomycin-Piperaquine (<i>Pf</i>), KAF 156 (<i>Pf</i>, <i>Pv</i>) , Cipargamin (<i>Pf</i> , <i>Pv</i>), Artefenomel (<i>Pf</i>), OZ439-Piperaquine (<i>Pf</i>), Tinidazole (<i>Pv</i>)
Phase III	Artemether spray (<i>Pf</i>), Azithromycin-Chloroquine (<i>Pf</i> , <i>Pv</i>), Pediatric Dihydroartemisinin-Piperaquine (<i>Pf</i>), Intra-rectal Artesunate (<i>Pf</i>, <i>Pv</i>) , Tefanoquine (<i>Pv</i>), OZ439-Ferroquine (<i>Pf</i>)

1.6 Problem statement and justification

Malaria remains a major public health concern with highest burdens in the tropical and subtropical regions. Approaches to combat malaria include vaccination, chemotherapy and vector control (biological, landscaping and use of insect treated nets). Although vaccination is the most effective disease protection strategy, a malaria vaccine has remained elusive¹³⁷. The European Medicines Agency approved Mosquirix™ as the first ever malaria vaccine in July 2015. However, the low protective efficacy (< 50%) and limited target group (children between 1.5–17 months) indicates it to be far from ideal. Annual fatalities from malaria are about 0.5 million in 2014¹. This drop compared to the previous years is due to the increased availability of ACTs and use of treated mosquito nets (TMNs)^{138,139}.

However, the gains attained so far in the containment of malaria infections and transmissions are greatly being hampered by the continued development of drug resistance by the *plasmodia* to virtually all antimalarial drugs designed. There is a growing fear that ACTs could be rendered ineffective as artemisinin resistance is spreading in Asia resulting in newer malarial infections and transmissions^{49,140,141}. Similarly the resistance to insecticides in the vector which is contributing to the disease resurgence¹⁴². Large pharma companies have lost their interest in research and development of newer antimalarial agents due to the cost of developing new drugs¹⁴³. Besides drug resistance and cost of production, toxicity resulting from undesirable side effects is a major public health concern. Moreover, the identification of effective drugs for all human groups is also a pressing need. Based on these factors, the search for new, cheap, effective and safe drugs to replenish the malarial drug pipeline remains a top priority. This could prevent the development of resistance as was with pyrimethamine-sulfadoxine (Fansidar®) in the 1980s.

From Table 1.2, majority of the available drugs and drug candidates undergoing testing are only effective against *Pf*. Importantly, for the successful elimination of malaria, novel drugs must have exclusive efficacy not only against the human *plasmodial* parasites but also the circulating wild types which may infect humans due to mutations. Therefore, careful identification of a drug target common to major *Plasmodium* species of interest is vital in achieving the elimination of malaria. Key metabolic pathways utilized by *plasmodia* for growth and replication are currently drawing intense research leading to the identification of molecular structures central to the functioning of this pathways. These include the haemoglobin degradation pathway and the subsequent heme group detoxification, fatty acid biosynthesis, oxidative stress and nucleic acid metabolism.

The genomes of vector, host and several *Plasmodium* species have been sequenced. This has allowed for the identification of drug targets which can be used in the R&D process. The *Pf*

genome contains ~100 proteases as potential antimalarial drug targets^{144,145}. One group of these proteases is the FPs, critical to *Plasmodia* life cycle mainly in the blood stage. Owing to the importance of these proteins in the development of malarial parasites, identification of small compounds that can modulate the activity of these proteases presents a unique window of opportunity to newer antimalarial drugs.

The advancement of computer technologies coupled with genomic data in recent times expedited the discovery and characterization of newer drugs^{89,146,147}. Combining bioinformatics in high throughput virtual screening allows millions of compounds to be screened yielding a small subset of hit compounds. The rich biodiversity of South Africa (SA) can be an important source of hit compounds against FPs and its homologs from other *Plasmodial* species. The recent development of SANCDB is expected to facilitate in the identification of potential hits for antimalarial drug discovery. Overall, the adoption of these computational approaches leading to the identification of novel chemotypes which might further be developed to potential antimalarial leads is important in the fight against malaria.

1.7 Hypothesis

Plasmodial cysteine proteases can be targeted with small non-peptide compounds leading to identification of potential antimalarial drug hits.

1.8 Research aim

The main aim of this study was to use computational approaches to determine the *in silico* antimalarial potency of selected compounds from the literature and identify related hits from South African natural sources with inhibitory activity against *Plasmodial* cysteine proteases. To achieve this, the study is subdivided into the following specific objectives:

- 1) Structurally compare FP-2 and FP-3 with homologs from other *plasmodial* species and human to determine the function features of these diverse enzymes (Chapter 2).
- 2) Docking of compounds from the literature, South African natural sources and “Zinc Is Not Commercial (ZINC)” into *plasmodial* and human proteins was performed by a previous MSc student. Molecular dynamic studies using GROMACS to determine the stability of these protein-ligand complexes. Additionally, establishment of a computational pipeline to perform molecular dynamic studies automatically to facilitate in the hit identification process (Chapter 3).
- 3) Binding free energy calculations between the protein-ligand complexes to quantitatively and qualitatively determine the strength and type of interactions involved and select the most promising hits (Chapter 4).
- 4) Use SANCDB, a chemical database recently established by RUBi to identify more chemical compounds which can be potential inhibitors against FPs and other *plasmodial* homologs (Chapter 5).

CHAPTER 2

Plasmodial cysteine proteases: In silico characterization

Haemoglobin degradation occurs exclusively in the digestive food vacuole, a digestive lysosome-like compartment of blood-stage plasmodia. Proteases involved include plasmepsins, histo-aspartic proteases and falcipains (FPs). In addition, the plasmodia invasion and rupture processes are characterised with abundance of proteolytic activity. This thesis focusses on FPs from Pf and other plasmodial species. Four FPs (FP-1, 2, 2' and 3) have been isolated allowing FP-2 and 3 to be validated as drug targets due to their developmental role for the parasites in the host. To facilitate the development of chemotherapies against these proteases, a better understanding of their structure, distribution and physicochemical properties is critical. This chapter focuses on the identification of FP-2 and FP-3 homologs from other plasmodial species and understanding their primary and tertiary structure besides analysing their physicochemical properties using in silico approaches. Human (host) cathepsins-L like proteases, homologous to plasmodial proteases, were also identified and compared in sequence, structure and biochemistry to identify differences exploitable for drug selectivity.

2.1 Proteases

Proteases (peptide hydrolases) are enzymes that hydrolyse peptide bonds of proteins and polypeptides. Proteases are continuously and universally produced all organisms. Depending on their specificity, these molecular scissors may target protein termini (exopeptidases) or cut within the peptide (endopeptidases). Proteases are grouped based on the principal catalytic residue: glutamate, serine, threonine, aspartate, cysteine and mixed^{148,149}. Each group is subdivided into clans and families. Proteases catalyze a wide array of biological reactions ranging from those involved in metabolic homeostasis to disease pathogenicity.

Genetic analysis has identified proteases as drivers of pathogenicity in parasitic diseases¹⁵⁰. The genome of *P. falciparum* codes for ~100 putative proteases of which ~30 are cysteine proteases¹⁴⁵. Pharmaceutical research initiatives have targeted this group^{151,152}. In *Plasmodium*, cysteine proteases activate pro-enzymes, degrade haemoglobin and participate in immunoevasion, tissue and cellular invasion as well as excystment^{148,153–155}. This functional diversity is due to the nucleophilicity of the cysteine, their stability and wide substrate range¹⁴⁸. Proteases specifically position the substrate on to the active site where the binding efficiency relies on the chemical environment of the subsite and the nature of the portion of the substrate peptide interacting directly with the active site groove¹⁴⁸. Despite the fact that during catalysis process only one peptide is hydrolyzed, a number of aa residues adjacent to the site of cleavage are crucial in determining the specificity and register of an enzyme¹⁴⁸. In cysteine proteases, a catalytic cysteine acts as the nucleophile. The additional electron shell of the Cys sulphur atom improves its nucleophilicity. The active Cys is deprotonated by an adjacent histidine. The resulting thiolate-imidazolium diad is often stabilized by an asparagine¹⁴⁸ (Figure 2.1). A glutamine creates an oxyanion hole to stabilize the tetrahedral intermediate. Cysteine proteases have a broad pH range due to the thiolate-imidazolium diad with a pKa of 4.0 for Cys and 8.5 for His.

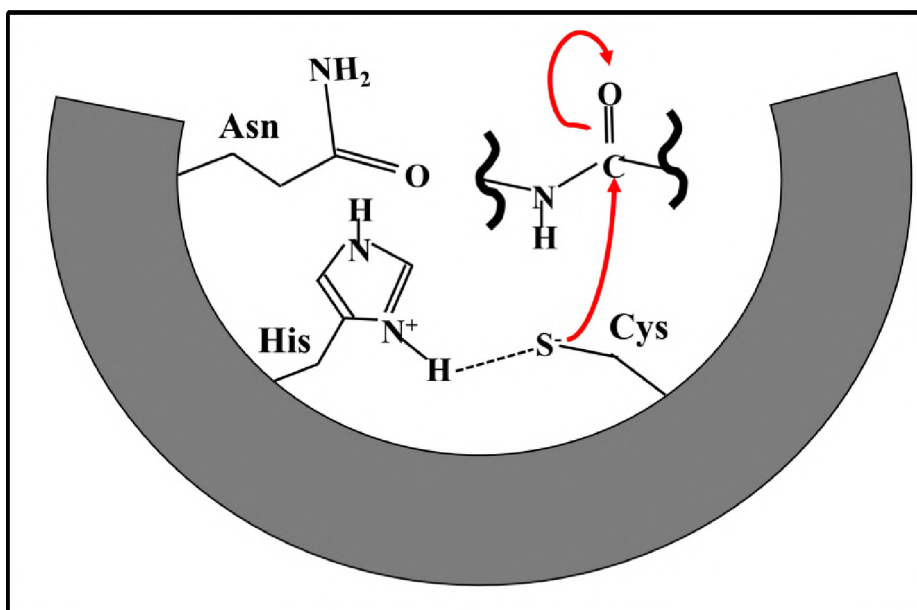


Figure 2.1: Catalytic mechanism of cysteine proteases. A schematic representation of the flow of electrons (arrows) and participating residues cleaving a peptide bond. Cys is the catalytic residue positioned near a proton withdrawing His group. Adapted from Erez *et al.*, 2009¹⁵⁶.

2.1.1 Cysteine protease nomenclature

Cysteine proteases are diverse with respect to sequence and structure. They are grouped according to their catalytic mechanism accounting for distinct evolutionary origins¹⁵⁷. They are divided into clans which are further divided into families using sequence homology, inserted loops and structural similarity¹⁵⁸. The MEROPS database of peptidases¹⁵⁹ lists 110 different cysteine protease families in 14 clans (Figure 2.2).

Groups most relevant to drug development are clans CA, CD and CE¹¹⁹. The papain and related human cathepsins of clan CA (papain-like proteases)¹⁶⁰ have been studied most intensively. Ten human cathepsins have been identified namely Cat B, C, F, H, K, L (L1 and L2) O, S, W and X or Z. They are expressed ubiquitously with Cat L participating mainly in the intracellular protein turnover. Clan CA proteases have a Cys-His-Asn triad conserved in their primary structure. The CA clan of *Pf* has four FPs, three dipeptidyl peptidases, nine serine-rich antigen (SERA) related proteins and a calpain homolog^{119,144}. This thesis focusses

on FPs, including FP-1, FP-2, FP-2' and FP-3¹¹⁹. Clan CD with a catalytic His-Cys dyad includes caspases and analysis of their sequences suggests that members of the C13 and C14 families are also present in *Plasmodium*. Clan CE proteases contain catalytic residues His, Glu (or Asp) and Cys in this order in the primary structure.

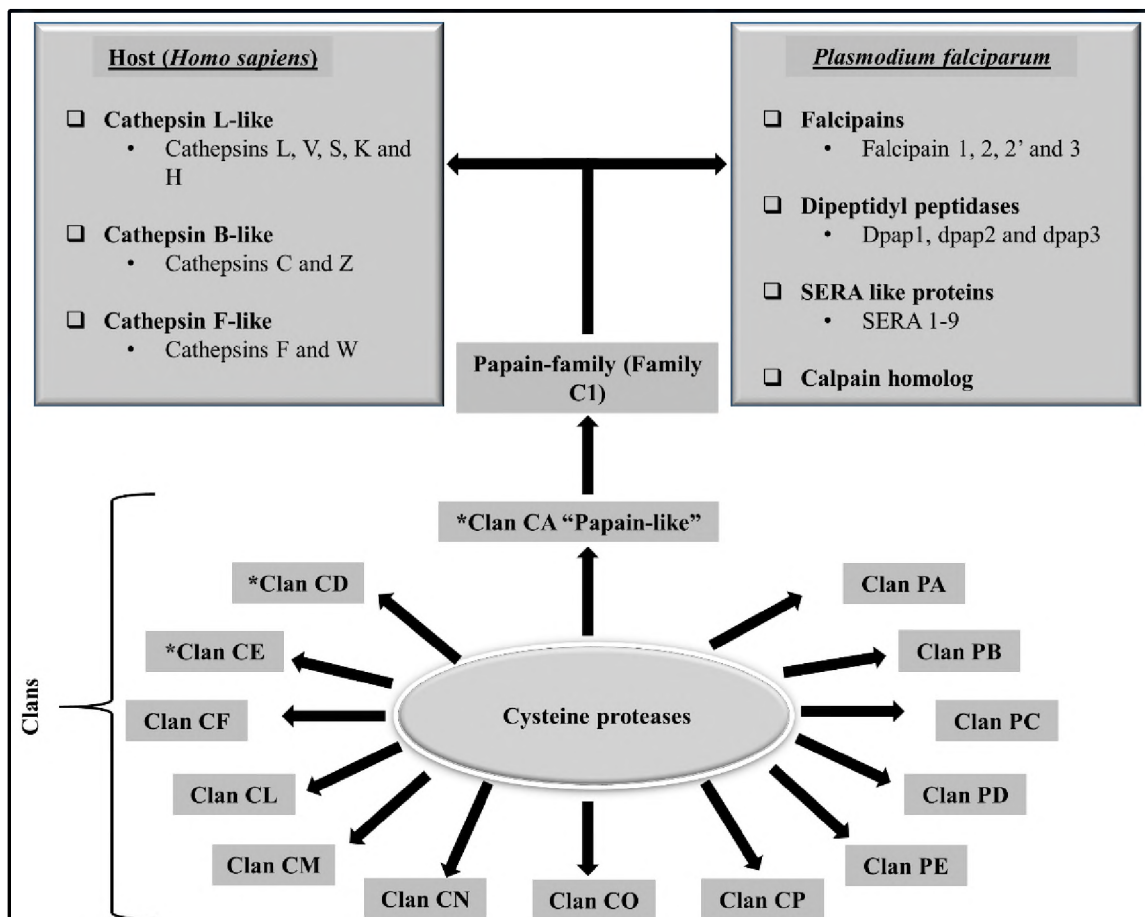


Figure 2.2: Nomenclature of cysteine (thiol) proteases. Clans with Cys as their catalytic nucleophile begin with a "C", mixed nucleophiles a "P". The second letter is sequential. Several cysteine proteases have not yet been classified. Clans of medical importance research are marked by an asterisk. Human and *plasmodial* papain-family proteases are also compared in the upper boxes.

2.2 Roles of *plasmodial* cysteine proteases

Plasmodial proteases production is strictly regulated with respect to time and location. Proteases are further controlled by endogenous inhibitors¹⁶¹. Cysteine proteases are essential for the survival and multiplication of the parasite making them prime drug targets for novel

antimalarial therapies¹⁶². Known protease inhibitors have provided much functional information for *plasmodial* cysteine proteases¹¹⁹. FP-2 gene disruption studies and timed additions of inhibitors Leupeptin and E-64 have confirmed the essentiality of FP-2 and FP-3^{97,119,163,164}. They function in parasite invasion and egression from the host cell, haemoglobin degradation, and intracellular development (Figure 2.3).

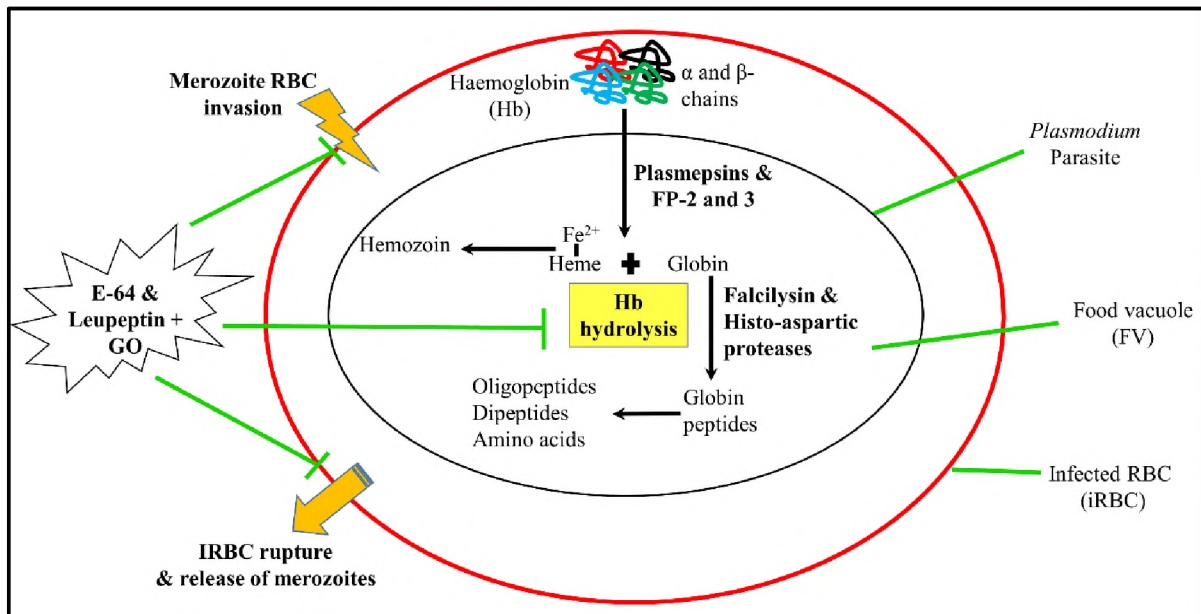


Figure 2.3: The role of cysteine proteases in the *Plasmodium* erythrocytic stage. Chemical inhibition and gene disruption studies identified proteases involved in haemoglobin degradation. FP mediated processes are shown in orange.

2.2.1 Haemoglobin hydrolysis

Plasmodium parasites lack aa synthetic pathway, and are solely dependent on host aa for growth and development. Merozoites invading erythrocytes consume more than half of the haemoglobin. The protein is taken up through a specialized organelle, the cytostome, and degraded in an acidic lysosome-like food vacuole by an array of proteases¹⁶⁵. Like lysosomes the food vacuole contains lysosomal cysteine proteases Cat B, H, L, K, S.

The globin is hydrolyzed to its constituent aa^{166,167} and cysteine proteases are central to the process²⁴. *Plasmodium* parasites treated with cysteine protease inhibitors E-64 and Leupeptin

successfully transport the erythrocyte cytosol to the food vacuole but are then unable to degrade the haemoglobin¹²¹. While several proteases participate in haemoglobin hydrolysis, the specific role of each enzyme and the sequence of events remains to be clarified¹⁶². Based on *in vitro* studies, several theories of haemoglobin degradation have been postulated. One proposed mechanism, aspartic proteases plasmepsin I and plasmepsin II were thought to initiate haemoglobin hydrolysis through degradation of the peptide bonds in the main chain residues of native haemoglobin releasing the heme moiety and the globin component which is further degraded by FPs^{119,166}. However, this explanation is faced by several limitations as it has been found that only cysteine protease inhibitors cause the swelling of the food vacuole indicating blocked haemoglobin degradation¹⁶⁸. *In vivo* studies concur that cysteine proteases initiate haemoglobin hydrolysis¹⁶⁹.

2.2.2 Tissue and erythrocyte invasion

During pre and erythrocytic stages, *Plasmodium* only grows and replicate within host cell such as hepatocytes and RBCs. For coordinated transmigration through different host tissues and cellular membranes *plasmodial* parasites require lytic enzymes to degrade the cytoskeleton^{15,170-173}. In hepatocytes, cysteine proteases are highly regulated to prevent premature apoptosis of infected cells. *Pf* inhibitor of cysteine proteases (*Pf*ICP or falstatin)¹⁷⁴ and *Pb* inhibitor of cysteine proteases (*Pb*ICP)¹⁷⁵ expression in *Pf* and *Pb* respectively during pre-erythrocytic stage suggests that there exists cysteine protease activity. However, they do not affect erythrocyte invasion by merozoites while serine protease inhibitors do¹¹⁹. A specific FP-1 inhibitor does block invasion reviving the debate on the role of cysteine proteases in invasion⁵³. Overall the serine proteases dominate in erythrocyte invasion while cysteine protease participation is uncertain³⁹.

2.2.3 Erythrocyte rupture

The erythrocytic cycle terminates in the rupturing of the infected erythrocyte releasing numerous merozoites. Cysteine protease inhibitors prevent erythrocyte rupture¹¹⁹. Leupeptin treatment of parasite cultures resulted in the accumulation of mature schizonts^{120,176}. Similarly E-64 blocks the lysis of the parasitophorous vacuole membrane surrounding the intraerythrocytic parasite by mature schizonts, implicating cysteine proteases in merozoites release²². Antipain and Leupeptin also block erythrocyte membrane lysis¹⁷⁷. The proteases dipeptidyl peptidase 3 (DPAP3) and *Pf* serine protease subtilisin-like protease 1 (*Pf*SUB1) centrally regulate egression process. For merozoite release, cysteine proteases thus first degrade protein networks of the parasitophorous vacuole followed by those of the iRBC¹⁷⁷.

2.2.4 Immuno evasion

Plasmodial cysteine proteases may either degrade host immune molecules or interfere with cellular immune responses leading to immune evasion³⁰. *In vitro* data involving protozoan parasites show that cysteine proteases can interfere with the antigen presentation process. For example, in *Trypanosoma cruzi* (*Tc*) and *Entamoeba histolytica*, cysteine proteases degrade host antibodies³⁰. *Tc* also blocks macrophage activation by blocking the NF- κ B P65 pathway via cysteine protease¹⁷⁸.

2.2.5 Exo-erythrocytic parasite stages

Data on cysteine proteases participation in non-erythrocytic parasite stages is minimal. One example is the hydrolysis of gametocyte surface protein *Pfs230*. The cleavage is blocked by E-64 indicating cysteine protease involvement⁵⁸. In addition, deletion of the gene encoding FP-1⁵⁹ and E-64⁶⁰ treatment of sexual-stage parasites decreased oocyst production, suggesting a specific role of this protease.

2.3 FPs cysteine proteases

FPs cysteine proteases of *P. falciparum* are related to papain family enzymes in sequence and function¹⁵⁵. Compared to other papain members they have longer prodomains and a 14 aa insertion in the carboxyl terminus of the catalytic domain^{179–181}. They are critical for the completion of the parasite lifecycle¹⁸². Four known FPs: FP-1, FP-2, FP-2' and FP-3 have been characterized biochemically¹⁵⁷. The gene encoding FP-1 is located on chromosome 14 of the others on chromosome 11¹¹⁹. The FP-1 catalytic domain has a sequence identity of only 40% to other FPs and its function in *plasmodia* development remain unknown¹⁸³. FP-2 and FP-2' are identical except for three aa away from the binding pocket¹⁸⁴. FP-2 and FP-3 share 68% sequence identity and have similarly sized prodomains. FP-2 and FP-3 also share N-terminal extension of the catalytic domain absent in FP-1¹¹⁹. FPs thus fall into distinct sub-families, FP-1 and FP-2/FP-3¹¹⁹. *P. vivax* encodes a homolog of FP-1 that shares 72% sequence identity in its catalytic domain¹⁸⁵ and 60-70% with the three homologs of FP-2/3.¹⁸⁶ We have identified additional FP-2/3 homologs from other *Plasmodium* species. These will be reported in the results section of the current chapter.

2.3.1 FPs expression profiles in iRBCs

Plasmodial gene expression is tightly regulated^{187–189}. Of the four FPs, FP-2 and FP-3 appear to be the major hemoglobinas responsible for the degradation of haemoglobin in the food vacuole¹⁵⁵. FPs are produced at different times during the blood stage and can functionally compensate for each¹⁹⁰. Immunoblotting data shows that FP-1 is expressed throughout the erythrocytic cycle and is active during the invasive merozoite stage as determined by immunofluorescence microscopy¹⁹¹ while FP-2 is maximally expressed in early trophozoites and the late trophozoites for FP-3^{164,181}. FP-2 accounts for over 90% of the cysteine protease activity in trophozoites¹⁸⁰.

2.3.2 Biochemical characterization of FPs

FP-2 and FP-3 exhibit very similar but not identical biochemical features, a sign of potential differences in functions. Both have low pH optima consistent with activity in the acidic food vacuole¹¹⁹. The specificity of papain-family proteases is determined by the P₂ position in the binding pocket and the two proteases (FP-2 and FP-3) possess a Leu amino residue in this position explaining their preference to peptidyl substrates¹¹⁹. FP-2 and FP-3 are produced as zymogens in specific cellular compartments and are activated at different rates¹⁹². There is no observable biochemical difference between FP-2 and its near-identical relative FP-2'. However, they have different expression profiles and the knockout of FP-2, but not FP-2' has a distinct phenotype. Thus, FP-2 cannot functionally be replaced by FP-2' whose function still remain unknown¹¹⁹. The biochemical evaluation of FP-1 is hampered by its low heterologous expression levels^{193,194}.

2.3.3 The structure and functions of the different FPs domains

Many clan CA cysteine proteases, including of FPs, are synthesized zymogens with a prodomain and a mature or active domain (Figure 2.4)^{119,155}. The FP-2 proenzyme for example is sized as a single 484 residues polypeptide whose N-terminal 243 aa prodomain is proteolytically removed during its transport to the food vacuole via the endoplasmic reticulum-golgi system. The pro-region inhibits premature cleavage of the catalytic domain, assists protein folding and acts as a signal for intracellular targeting¹⁴⁸. Removal of prodomain releases the activated mature enzyme¹⁵⁵. The prodomain consists of several conserved motifs while the mature domain may be subdivided into a left (L) and a right (R) subdomain. This thesis focusses on the catalytic domain. The catalytic residues Cys42, His174 and Asn205 are found in a cleft between the L and R domains¹⁵⁵. FP-2 and FP-3 share a “nose-like” projection connecting the L and R sub domains and a C-terminal arm¹⁹⁵.

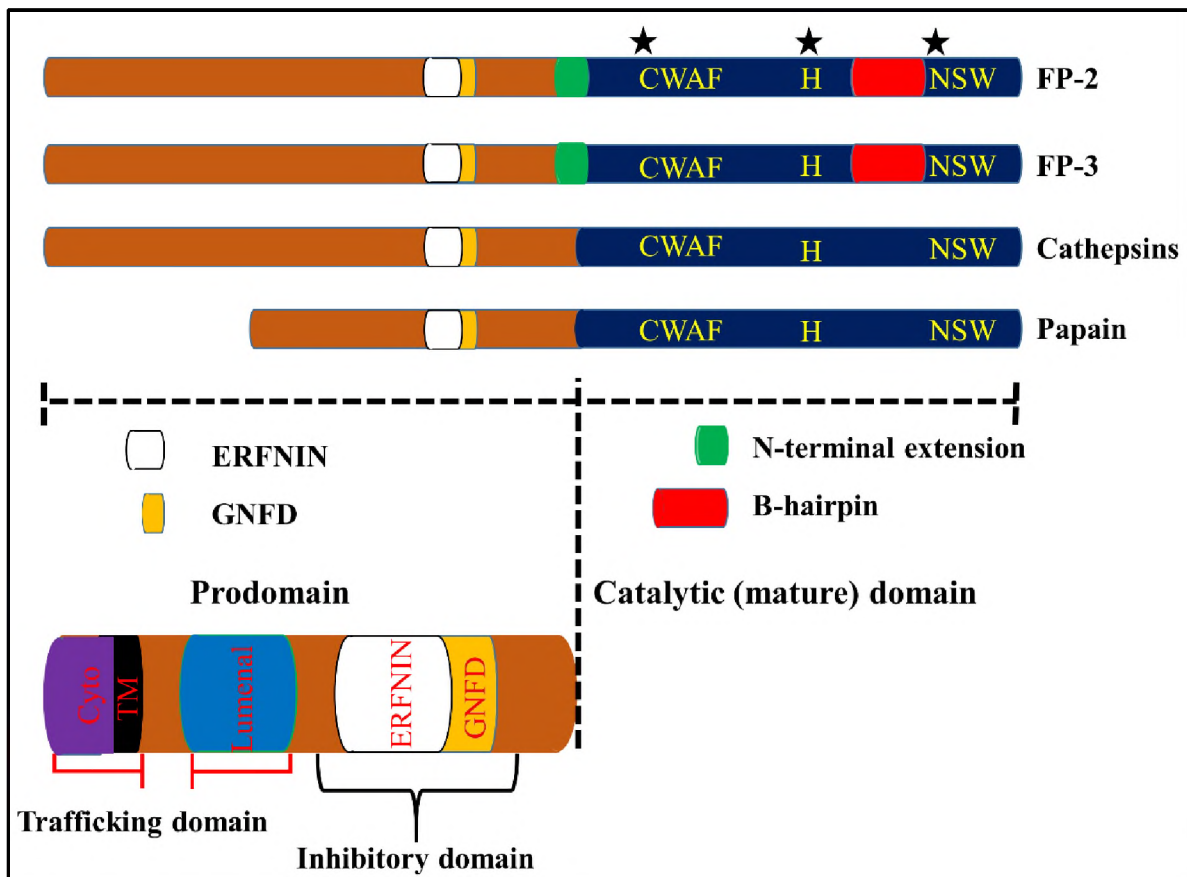


Figure 2.4: Papain proteases structure. Key structural elements and their position in FPs, cathepsins and papain (prototype). Key residues forming the catalytic triad are marked with an asterisk. Adapted from Pandey KC and Dixit R, 2012¹⁵⁵.

2.3.3.1 Falcipain prodomain

The prodomain controls the activity of the catalytic domain of FPs and homologs before activation and the N-terminal part is responsible for the trafficking of FPs into the food vacuole¹⁵⁵. The prodomain consists of a 35 aa cytoplasmic part, a 20 aa transmembrane α -helix and a 188 aa luminal part¹⁵⁵ (Figure 2.4). Green fluorescence protein (GFP) fusions of FP-2 and FP-3 lacking the transmembrane part of the prodomain localized to the cytoplasm. Serial truncation experiments implicated 20 aa of the luminal domain and 10 aa of the cytoplasmic domain in controlling proper localization¹⁵⁵. Gene constructs encoding different parts of the prodomain in combination with FP-2¹⁹⁶ indicated that the C-terminus of the prodomain endogenously inhibits FP-2 with two highly conserved motifs ERFNIN and GNFD particularly crucial. These motifs conserved in all FPs, their homologs and cathepsin

L subfamily proteases. Together with hydrophobic residues Phe and Trp, the two motifs further ensure the secondary structure of the prodomain is maintained. The secondary structure of the inhibitory domain (Leu¹⁵⁵-Asp²⁴³) is not retained when the prodomain lacks the sequence downstream of the ERFNIN and GNFD motifs or in a peptide fragment spanning the two motifs¹⁵⁵.

2.3.3.2 Mature domain

Successful trafficking of FPs and prodomain processing yields the mature domain encompassing residues 244-484 for FP-2 and 243-492¹⁹⁶ for FP-3. The catalytic domains of FP-2, FP-3 and some *plasmodial* homologs are the only papain-family proteases successfully re-folded under alkaline conditions without their prodomains^{197,198}. For correct folding of FP-2 *in vitro*, 17 residues preceding the N-terminus of the catalytic domain are required¹⁹⁹. Interestingly, this polypeptide mediates folding when covalently linked to the catalytic domain or when added to the buffer²⁰⁰. Despite function conservation of the FPs refolding domains, their sequence identity is only 20-45%. Refolding domains of other *plasmodial* proteases induce correct folding of FP-2 with similar kinetics as the wild-type peptide indicating that N-terminal extensions are functionally equivalent¹⁵⁵. FP-2 and FP-3 folding domains though a small part of the prodomains still adopt a defined secondary structure. Stabilization of the catalytic domain by the folding peptide involves a buried hydrogen bond between Tyr¹³ and Glu¹²⁰, and a salt bridge of Arg⁵ residue¹⁹⁵. FPs has a 14-residue insertion near its C-terminus^{201,202} sandwiched between the active site His and Asn. It helps to bind haemoglobin^{155,201} and is found in all *plasmodial* FPs with a weak aa conservation.

2.4 Structural basis of falcipain inhibition

Cysteine proteases are central to the erythrocytic parasite life cycle, and hence constitute potential drug targets for the development of novel antimalarial drugs¹⁵⁵. Both *in vitro* and *in vivo* studies confirm the antimalarial potency of protease inhibitors^{155,203}. FP-2 is the main

trophozoite cysteine protease and although its deletion is not lethal, it prevents haemoglobin cleavage, and increases parasite susceptibility to cysteine proteases inhibitors. FP-3 knock-out, by contrast, is lethal making it an important drug target. Structural analysis of FP-2 and FP-3 has demonstrated small molecules and protein inhibitors as leads in antimalarial drug discovery²⁰¹. Both FP-2 and FP-3 are validated drug targets²⁰⁴. Both peptide and non-peptide inhibitors of pathogen proteases are possible whether reversible or irreversible¹⁶². Most FP-2 and FP-3 inhibitors are peptidic¹⁸⁴ and include α -ketoamides²⁰⁵, E-64 epoxysuccinyl derivatives²⁰⁶ and peptidyl aldehydes²⁰⁷. However, non-peptidic compounds could overcome the degradation of peptide based inhibitors. Crystal structures of FP-2 in complex with cystatin and E-64 as well as FP-3 with Leupeptin and K11017, a vinyl sulfone inhibitor have indicated this to be possible^{163,208}. These structures identified an FP_{nose} and an FP_{arm} involved in the protease folding and haemoglobin interaction respectively. A crystal structure of the falcipain-haemoglobin complex is not yet available. Computer simulations for drug development is becoming increasingly important²⁰⁹. Thus, it is of interest to utilize both ligand and structure guided drug design methods to probe the structural requirements of potent falcipain inhibitors which may provide lead molecule to obtaining novel efficacious antimalarials²¹⁰.

2.5 Proposed work

FP-2, FP-3 as well as *plasmodial* and human homologs will be compared in structure and sequence using phylogenetics, sequence alignment, physicochemical properties and motif signatures to identify any significant differences.

2.6 Methodology

Several web based bioinformatics tools and databases were used for analysis or data source (Appendix 1A). Figure 2.5 summarizes the workflow the *in silico* approach of this chapter.

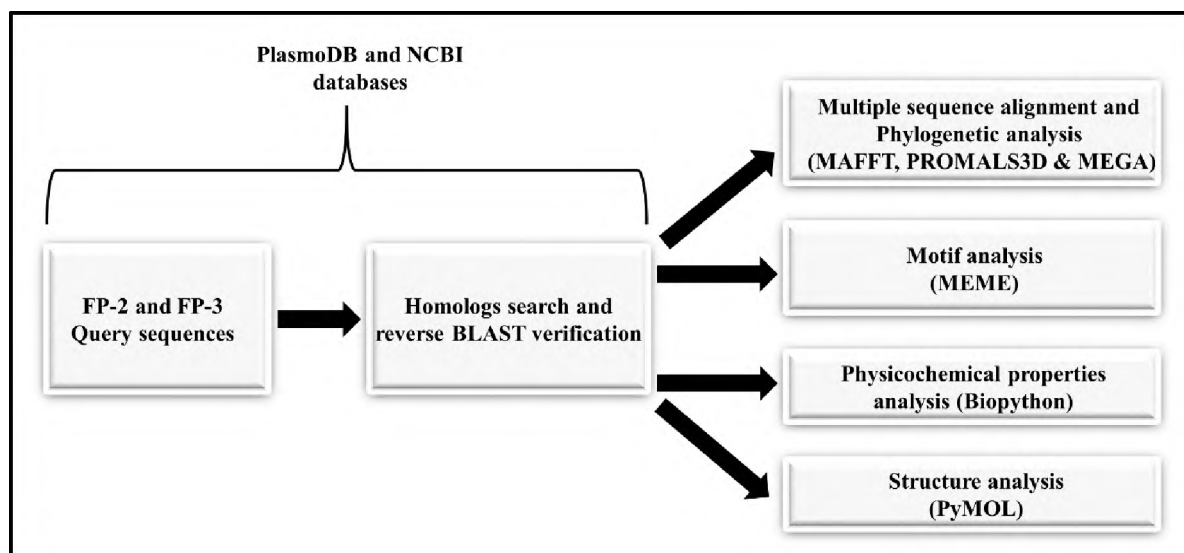


Figure 2.5: Analytic approaches applied to FP-2, FP-3 and homologs.

2.6.1 Data acquisition

2.6.1.1 Protein aa sequence retrieval

FP-2 and FP-3 protein sequences (accession numbers: PF3D7_1115700 and PF3D7_11154400) were retrieved from PlasmoDB²¹¹. Using BLASTP, these sequences were used to identify other *plasmodial* and human homologs in PlasmoDB and NCBI²¹² databases using default parameters. A reverse BLAST²¹³ search was performed to ascertain if the retrieved hits were the true orthologs. Sequences with E-value below 10^{-5} and with significant query coverage were selected (Appendix 1B).

2.6.1.2 Retrieval of 3D protein structures

Coordinate files for the crystal structures of FPs (FP-2 [PDB ID: 2OUL] FP-3 [PDB ID: 3BWK]) and human cathepsins (Cat K [PDB ID: 3OVZ], Cat L [PDB ID: 3OF8 and Cat S [PDB ID: 1NPZ]) were retrieved from the Protein Data Bank (PDB)²¹⁴. For other *plasmodial*

proteases, homology models previously prepared by an MSc student²¹⁵ using MODELLER version 9.10²¹⁶ were used.

2.6.2 Multiple sequence alignment and subsite composition analysis

Multiple sequence alignment (MSA) of FP-2, FP-3 and other *plasmodial* homologs were performed using online servers MAFFT²¹⁷ and PROMALS3D²¹⁸ (Appendix 1C). As alignments of sequences are inherently uncertain²¹⁹, MSA outputs were compared to determine their alignment accuracy. For MAFFT, the following sequence alignment parameters were used; BLOSUM62 scoring matrix was used as substitution matrix with a gap opening and extension penalty of 1.53 and 0.123. All other parameters were set as default. The number of tree building steps was set to 2 with a maximum iteration set of 2. For PROMALS3D, default parameters were used with an exception of PSI-BLAST expect value which was adjusted to 0.0001. FP-2 [2OUL] and FP-3 3D [3BWK] structures were used to add constraints to the alignment. The catalytic domains of the proteases were obtained by trimming the prodomains using JalView software^{220,221}. The remaining mature (catalytic) domains were realigned to increase accuracy. JalView tools were used to determine the sequence identities, similarities and visualize conserved regions. To determine subsite aa conservation and variation, all subsite residues from each protease were extracted from the alignment into a Fasta file and visualized using WebLogo webserver²²².

2.6.3 Phylogenetic analysis

Determining the best substitution models for evolutionary inference is the first step for correct phylogenetic analysis. Using MEGA5.2 (Molecular Evolutionary Genetic Analysis)²²³ software the evolutionary relationship the catalytic domains of all proteases (*plasmodial* and human cathepsins) was determined. The following analysis preferences were set; Neighbor-joining tree, statistical method (Maximum Likelihood), and substitution model (aa substitution type). Up to 48 aa substitution models were calculated for both the complete

(100%) and partial (95%) deletion. The best three models, selected according to lowest BIC scores are shown in Appendix 1D. Phylogenetic tree construction for each of the selected models was performed and compared to determine the robustness of the tree construction process. Gamma (G) evolutionary distance correction value was set as determined for each of the models and Nearest-Neighbor Interchange (NNI) chosen as the tree inference method. For each model, complete or partial deletion method of gap (missing data) treatment and bootstrap value of 1000 was used for tree construction.

2.6.4 Motif discovery

To identify the existence and distribution of motifs within FP-2, FP-3 and their homologs, a standalone MEME suite (version 4.10.2) was used under Linux operating system. On the command line, a Fasta file containing all the sequences was parsed to the MEME analysis software with the following analysis preferences; -nostatus -time 18000 -maxsize 160000 -mod zoops -nmotifs X -minw 6 -maxw 50. Where -nmotifs X was varied until no more motifs were discoverable. Using an in-house Python script *viz. motif_analyzer.py* (Appendix 2A), a heatmap representing the distribution of the different motifs was generated. Using PyMOL (The PyMOL Molecular Graphics System, Version 1.6.0.0 Schrodinger, LLC.), the different motifs were mapped onto the protein structures.

2.6.5 Physicochemical properties

To determine the aa composition and physicochemical properties *viz.* molecular weight (Mr), pI (theoretical isoelectric point), aromaticity, instability index, aliphatic index and GRAVY (Grand Average of Hydropathy) of all key proteases, an *ad hoc* Python script *viz. props.py* (Appendix 2B) utilizing the Biopython module was used. Appendix 1E is the three and one letter annotation of all residues.

2.7 Results and Discussion

2.7.1 Sequence analysis

2.7.1.1 Protein sequences

Using the query sequences of FP-2 and FP-3, up to 13 *plasmodial* and three human homologs were retrieved from the PlasmoDB and NCBI databases. From the reverse BLAST results, all the retrieved sequences were FP-2 and FP-3 true orthologs. Interestingly, in most reverse BLAST results, the first hit was to FP-3 even when FP-2 was used as the query sequence indicating inaccurate annotation of these homologs in literature. Table 2.1 shows FP-2 and FP-3 identified homologs from significant *Plasmodium* species and human cathepsins and their corresponding sequence identities. For a complete list of all identified homologs and their corresponding source organism (Appendix 1B).

Table 2.1: Key FP-2 and FP-3 homologs from different *plasmodial* proteases. Adapted from Musyoka TM *et al.*, 2015²²⁴.

Accession number	Common name (Abbreviation)	Source organism (Abbreviation)	% SI	
			FP-2	FP-3
PF3D7_1115700	Falcipain-2 (FP-2) [§]	<i>P. falciparum</i> (Pf)	100	66
PF3D7_1115400	Falcipain-3 (FP-3) [§]		66	100
PVX_091415	Vivapain-2 (VP-2)	<i>P. vivax</i> (Pv)	62	66
PVX_091410	Vivapain-3 (VP-3)		57	57
PCHAS_091190	Chaubaudipain-2* (CP-2)	<i>P. chabaudi</i> (Pc)	50	48
PKH_091250	Knowlesipain-2* (KP-2)	<i>P. knowlesi</i> (Pk)	57	57
PVX-091260	Knowlesipain-3* (KP-3)		57	60
PBANKA_093240	Berghepain-2* (BP-2)	<i>P. berghei</i> (Pb)	51	47
PY00783	Yoelipain-2* (YP-2)	<i>P. yoelii</i> (Py)	48	47
gi 157830076	Cathepsin-K (Cat K)		38	41
gi 313754424	Cathepsin-L (Cat L)	<i>H. sapiens</i>	37	38
gi 30749675	Cathepsin-S (Cat S)		36	37

* = adopted names for convenience

2.7.1.2 Multiple sequence analysis

Table 2.2 shows the position of the catalytic domain in the whole length of the corresponding protein as well as numbering adopted in the thesis (for convenience purposes). After minor manual adjustments, MAFFT MSA output was considered to be the best as it aligned key residues correctly. The query-hit sequence identities and residue conservation were determined (Table 2.1 and Figure 2.6a).

Table 2.2: The position of the catalytic domain within the whole protein sequences of different FP-2 and FP-3 homologs. Adapted from Musyoka TM *et al.*, 2015²²⁴.

Protein	Position in whole sequence	Mature domain numbering
FP-2	244-484	1-243
FP-3	250-492	1-249
VP-2	246-487	1-242
VP-3	253-493	1-241
KP-2	252-495	1-244
KP-3	240-479	1-240
CP-2	231-471	1-241
BP-2	228-468	1-241
YP-2	232-472	1-241
Cat K	115-329	1-215
Cat L	113-333	1-221
Cat S	115-331	1-217

MSA results identified that up to 45 aa (highlighted in green) which included the clan C1A characteristic catalytic triad residues namely Cys, His and Asn (marked with an asterisk) and the Gly-Cys-X-Gly-Gly motif were fully conserved in all protein sequences. Up to 18 aa positions were conserved in the *plasmodial* proteases only (highlighted in blue) while the human cathepsins had 23 unique aa (highlighted in black). The rodent *plasmodial* homologs had 34 aa positions exclusively conserved (highlighted in grey). On the other hand, the human *plasmodial* proteases had only 4 unique aa positions (highlighted in red).

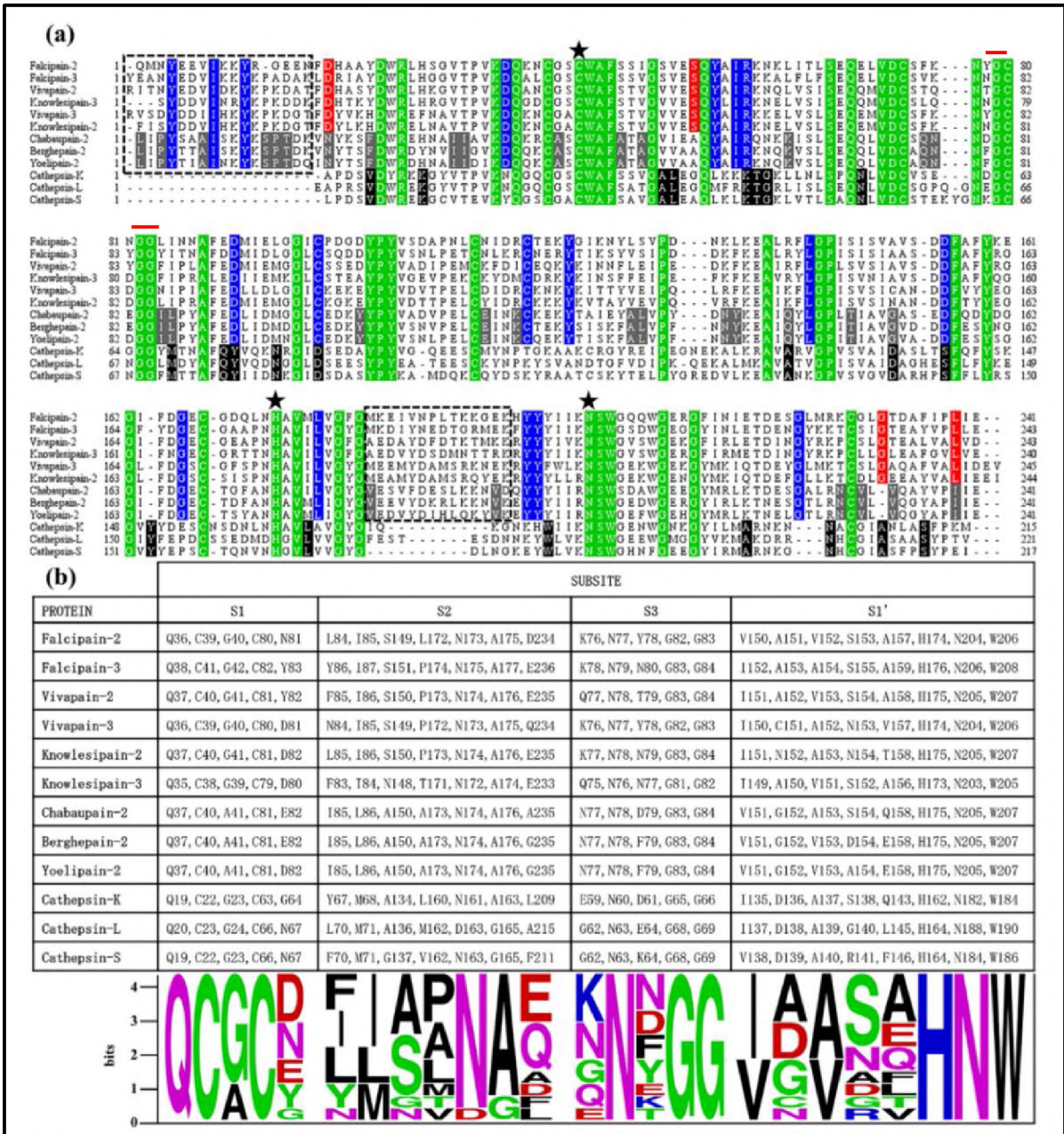


Figure 2.6: Sequence analysis. a) MSA output from MAFFT alignment software. Marked with an asterisk is the characteristic catalytic triad residues while the solid red line shows the GCXGG motif. Highlighted in green are aa positions that 100 % conserved in all sequences, blue, black, red and grey showing only aa positions exclusively conserved in *plasmodial*, human cathepsins, human and rodent *plasmodial* homologues in that order. The two inserts characteristic of the *plasmodial* proteases are enclosed with a broken line b) Individual subsite composition (table) and corresponding aa conservation (weblogo). From Musyoka TM *et al.*, 2015²²⁴.

As has been the case with FP-2 and FP-3, two aa inserts were also present only in the *plasmodial* proteases (boxed). The first commonly referred to as the nose consists of

approximately 17 aa occurred at the N-terminus while the second, the arm was near the C-terminus. Pandey *et al.*, previously established that the nose was responsible for the correct folding of the catalytic domains in FP-2 and FP-3¹⁵⁵. The arm which consists of approximately 14 aa forms a highly flexible β -hairpin. This has been linked to the haemoglobin (substrate) binding²⁰⁸. The arm aa composition is fairly conserved within the FPs with significant variations being observed in other *plasmodial* orthologues. As determined in FP-2 and FP-3, the functions of the inserts may play identical roles in the other *plasmodial* homologs although this has to be determined. Notable was the varied aa composition between the human *plasmodial* homologs and the rodent counterparts.

Disulphide bonds play critical role in the determination of the overall 3D structure of proteins formed through the ER pathway²²⁵. Besides the main catalytic Cys centre, up to 8 other Cys residues were observed in the *plasmodial* proteases and participate in the formation of 4 disulphide linkages. In contrast, the human homologs have only 3 pair of disulphide cysteine forming residues.

2.7.1.3 Structure and composition of the binding pocket of cysteine proteases

Determining the structure of an enzyme as well as understanding its specificity are indispensable for structure-based drug design²²⁶. Thus, evaluating aa composition, shape, size and volume is critical in designing inhibitors with selective activity for *plasmodial* proteases. The binding pocket of cysteine proteases is situated in a cleft between the structurally conserved R and L domains. The aa surrounding the catalytic Cys residue are organized into 4 subsite *viz.* S1, S2, S3 and S1'¹⁶³. It is of paramount importance to determine the differential aa composition of the active site of the human cathepsins and *plasmodial* cysteine proteases. This establishes important aa differences that can be targeted for attaining drug selectivity. From the literature, residues forming the subsites of FP-2, FP-3, and VP-2 have been established. From the alignment, subsite residues of other homologs have been

identified by mapping those of the known proteases. WebLogo analysis results revealed that the S1 and S3 subsite residues were fairly conserved across all the *plasmodial* proteases. However, S2 and a portion of S1' are highly varied (Figure 2.6b). In all proteases, aa residues polarizing His174 (S1') during hydrolysis namely Gln36 (S1), Asn173 (S2) and Asn204 (S1') are conserved with the exception of the S2 position of Cat L which has an Asp residue (numbering as per FP-2 catalytic domain). The two conserved cysteine residues in S1 namely Cys39 and Cys80 form a disulphide bridge critical in stabilizing the proteins²⁰¹. Another residue conserved in all proteases except rodent proteases is Gly40 residue. In the rodent homologs it is replaced by Ala. The fifth position of S1 is highly variable although all residues were all polar suggesting a conserved function. S2, the major pocket determining ligand specificity in cysteine proteases is mainly hydrophobic¹⁵⁵. The hollow (deepest) end of S2 in human *plasmodial* homologs has a polar charged residue. In contrast, the rest of homologs including the cathepsins have a small uncharged residue at the same position. S2 of human cathepsins was composed entirely of hydrophobic residues²²⁷.

For structure guided drug design is the volume of the active site is critical as this determines the ligand specificity. The S2 pocket volume differs between FP-2 and FP-3 as well as VP-2 and VP-3. In FP-2, the S2 opening groove is formed by smaller Leu84 and Leu172 whereas in FP-3, more bulkier Tyr86 and Pro174 exist narrowing the distal end²²⁸. For VP-2, the S2 is broader than in VP-3 as the region in between S1' and S2 subsites of VP-3 is hanging inwards making its S2 narrower²²⁹. The location of the subsites in cathepsins and *plasmodial* homologues is shown in Figure 2.7.

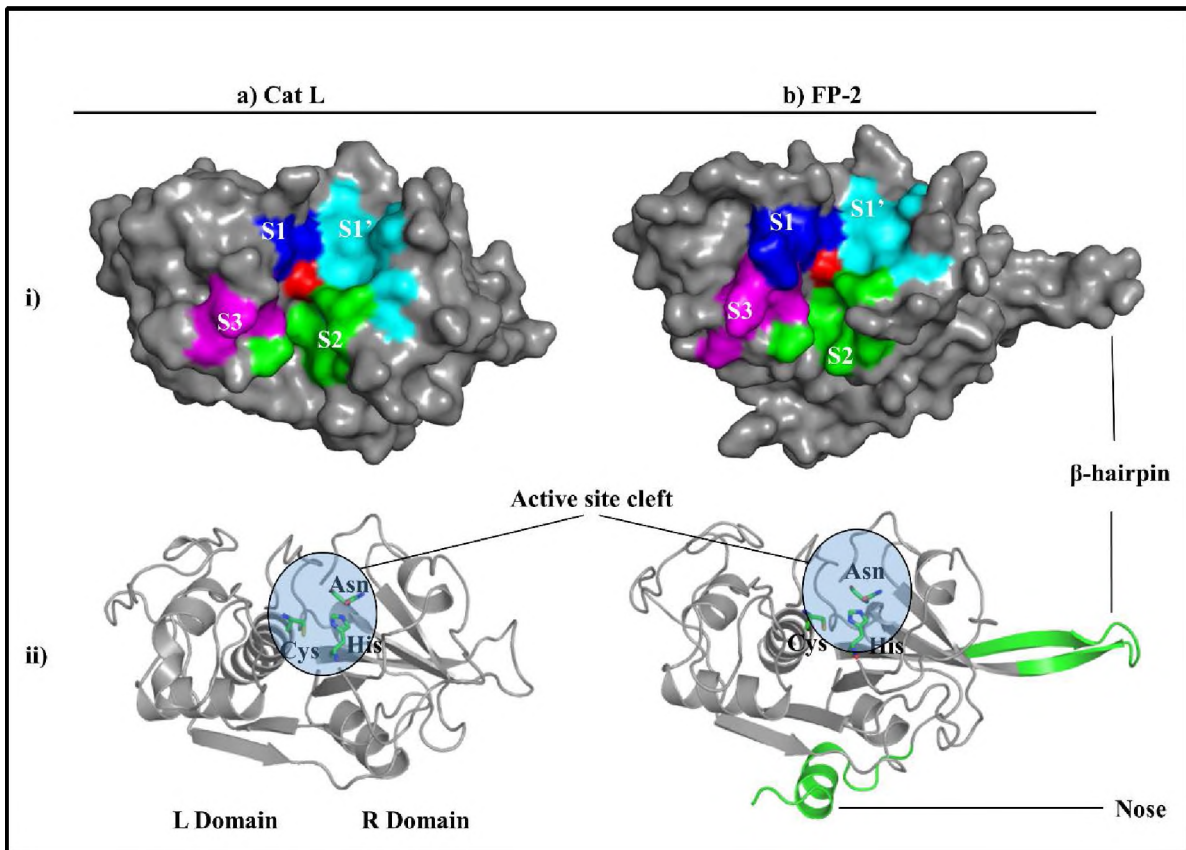


Figure 2.7: The structure of the cathepsins and FPs. A surface representation i) and cartoon representation ii) of Cat L (a) and FP-2 (b). The different subsites that form the binding pocket of the proteases (i) are highlighted as blue (S1), green (S2), S3 (magenta) and S1' (magenta). Marked with red is the catalytic Cys residue. Clan C1 enzymes structural fold consists of the left and right domain (ii). The unique features (nose and B hairpin) that are characteristic of the FPs and other *plasmodial* homologues are highlighted in green (bii).

The S2 opening of rodent *plasmodial* homologs has highly conserved less bulky Ile85 and Ala173 aa on either side. In all proteases, S3 has a highly conserved Gly-rich component which is critical in the stabilization of substrates via hydrogen bonding²²⁹. A highly conserved Trp residue is positioned at the opening of the cleft in all the protease and participates in the formation of hydrogen bonding with substrates²²⁸.

2.7.2 Phylogenetic analysis

Plasmodial proteases and human cathepsins share a common proteolytic mechanism with other papain-like cysteine proteases²³⁰. To infer evolutionary differences between the two protease classes, both defined proteins (Figure 2.8) and all proteins (Appendix 1F) were

investigated phylogenetically. This may reveal evolutionary differences important in designing inhibitors exclusively targeting *plasmodial* proteases to achieve inhibitor selectivity.

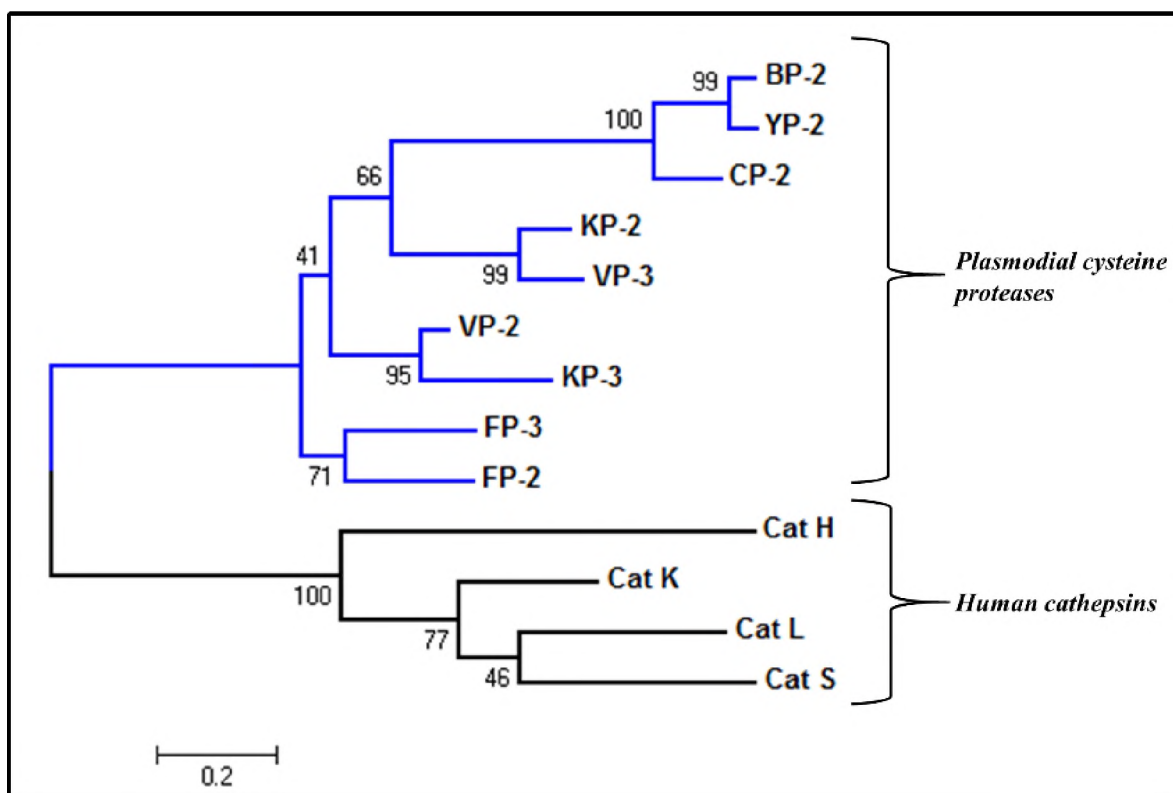


Figure 2.8: Evolution analysis. A phylogram of human and *plasmodial* FP-2 and FP-3 and homologs catalytic domain using MEGA5.2.2. The evolutionary relationship was inferred using a maximum likelihood version of the Whelan And Goldman (WAG) model²³¹. Initial trees for the heuristic search were obtained using the Neighbor-Joining method to a matrix of pairwise distances estimated using a JTT model. A discrete γ distribution (+G) parameter of 2 was used to model evolutionary rate differences among sites while the rate variation model allowed for some sites to be evolutionarily invariable ([+I], 6.5% sites). All positions containing gaps were completely eliminated. The level of bootstrap support was inferred by 1000 resampling of the alignment. Numbers on branches represent posterior probabilities as percentages (cut off > 50%). The scale bar represents the number of aa substitutions per site.

In the phylogenetic tree, the most notable observation was the distinct clustering of the human and *plasmodial* homologs into two discrete groups (Figure 2.8). A similar clustering of human cathepsins in phylogenetic studies of (Appendix 1F). This is consistent with the low sequence identity observed between FP-2 and human cathepsins (Table 2.1 and Appendix 1B). Rodent associated *plasmodial* enzymes also form a separate group from their

human *plasmodial* counterparts, a fact supported by the differences observed from the MSA (Figure 2.6a, Appendix 1F). The clustering is important as rodent models are used to develop human antimalarial drugs. Understanding the differences in inhibitor interaction of human *plasmodial* FPs is invaluable. Notably, Cat H with both exopeptidase and endopeptidase activity evolved away from Cat K, L and S with the endopeptidase activity only. A gene speciation event may have caused the separation of primate and murine enzymes. Interestingly, Cynomolgipain-2 (CyP-2), a FP-2 homolog appears the most evolved followed by and VP-2. The two species share phenotypic, biological and genetic characteristics compared to *Pk*²³². Among FPs, FP-2 and FP-3 form a distinct group possibly due to high sequence identity between the two. VP-2 and KP-3, form a separate group and closest to FP-2 and FP-3 cluster (Figure 2.8). Interestingly, human infecting *Pm* and *Po* proteases, Malariaepain-2 (MP-2) and Ovalepain-2 (OP-2) group with Gallinacepain-2 (GP-2) [birds] and Reichenopain-2 (RP-2) [Chimpanzee] (Appendix 1F).

2.7.3 Motif analysis

Protein sequence motifs are aa sequence patterns that often carry biological function in homologous proteins. They are frequently used to classify proteins. The programme MEME²³³ is widely used to find motifs in DNA or protein sequences. Finding motifs in a group of protein families and their occurrence can provide important insights into the protein functions. A coherent approach is required to identify unique motifs. The maximal number of unique motifs discoverable within our set of sequences was determined and ranked based on the MAST E-value (a product of number of sequences in a database and combined position *p*-value of each sequence)²³³. Here proteases listed in Table 2.1 were used. Up to 9 unique motifs of different lengths (max width = 50 aa, min width = 8 aa) were identified (Appendix 1G). These motifs M1, M3 and M5 were present in both *plasmodial* and human proteases and indicate a conserved functionality (Figure 2.9). Human cathepsins have two (M7 and M8)

and human and rodent *plasmodial* proteases four unique motifs (M2, M4, M6 and M9) although FP-2 lacks M6.

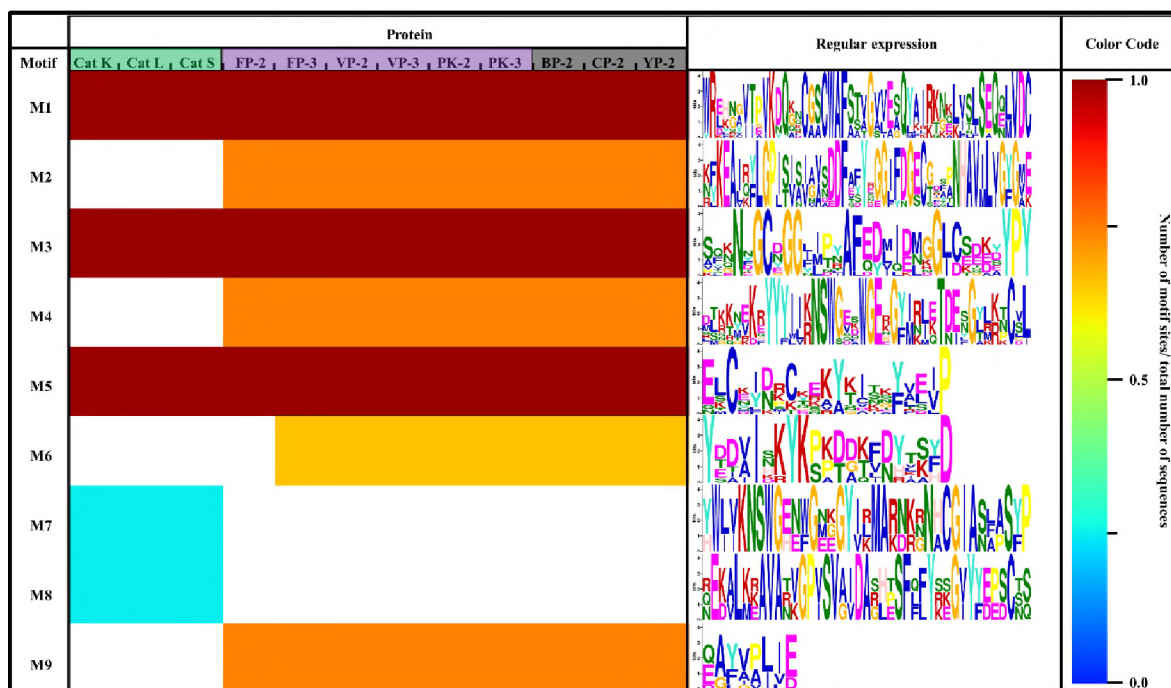


Figure 2.9: Motif analysis. A heatmap and “regular expressions” of motifs of FP-2 and FP-3 and their homologs (*plasmodial* and human) identified by MEME software. Proteins lacking the named motif are shown in white. The color code shows the level of motif conservation with red and blue being the highest and lowest respectively.

To better understand the function of the identified motifs, they were mapped onto the respective protein structures using PyMOL (Figure 2.10). Interestingly, widely conserved motifs M1, M3 and M5 mapped to the L-domain while the R-domain harbours motifs that differentiate the two groups of proteases (human and *plasmodial* proteases). Using the best possible match of regular expression (Appendix 1G), the PROSITE webserver was used to determine the functional importance of identified motifs. M1 encompasses three functional sites PS00139 (QQnCGSCWafST) with the catalytic cysteine, PS00008 (GScwAF or GVvesSQ) with N-myristoyl and PS00006 with a casein kinase II phosphorylation site (SslE). M3 contains an N-myristoylation site and protein kinase C phosphorylation site. M5 includes two phosphorylation sites PS00005 for Protein kinase C and PS00004 for cAMP or cGMP-dependent protein kinase plus N-myristoylation site.

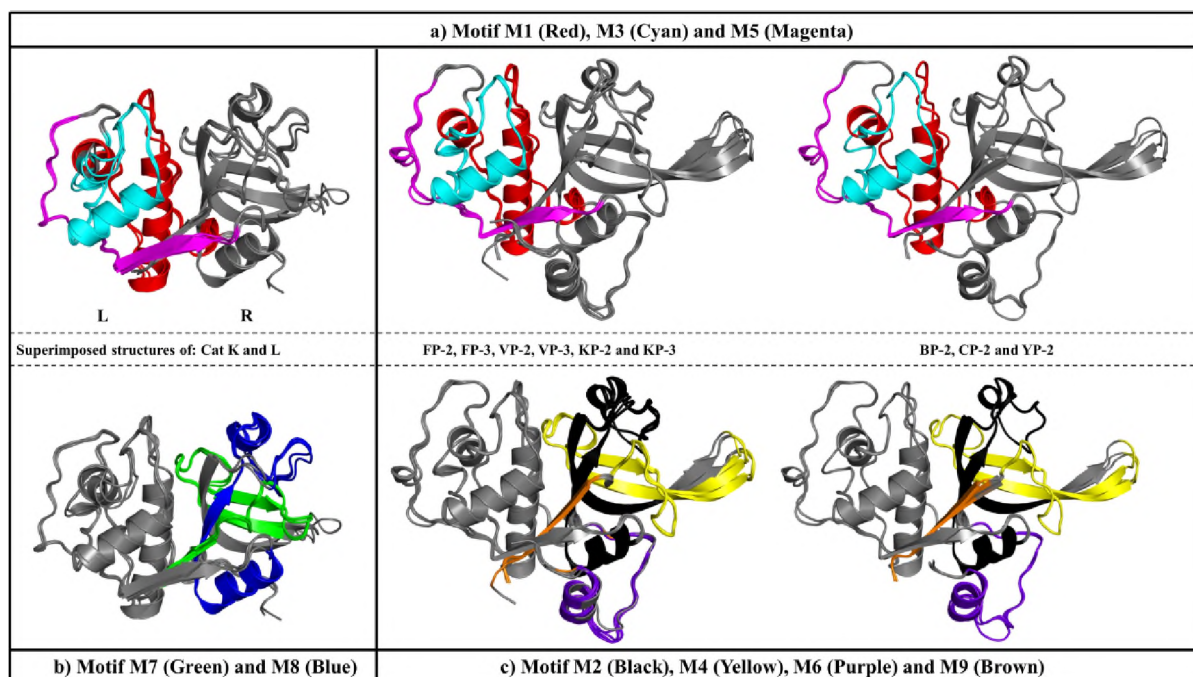


Figure 2.10: Location of motifs. The location of identified sequence motifs the structures of FP-2 and FP-3 and their homologs (*plasmodial* and human). Motifs concerned in: a) all proteins, b) human cathepsins only and c) *plasmodial* proteases.

For M2 and M6, only N-myristoylation site and a casein kinase II phosphorylation site were found respectively, while no modification site was detected for M4. For the motifs unique to human cathepsins M7, bears a thiol protease asparagine active site (PS00640) and a casein kinase II phosphorylation site while no functional site was detected for M8. Protein phosphorylation is key in protein regulation through a series of signalling cascades. In *Pf*, a kinome of ~99 protein kinases regulate key steps affecting all the stages of the parasite cycle²³⁴. However, the relevance of the identified phosphorylation sites within the FPs and their homologs remains unclear. N-myristoylation of FPs has not been investigated. However, N-myristoyltransferase (NMT) is a validated implying direct relevance²³⁵. In other apicomplexan parasites such as *Trypanosoma brucei* (*Tb*) and *Tc* and *Leishmania*, the process of myristoylation is critical for parasite growth. NMT inhibitors correspondingly allowed clearing *Tb* parasites from infected rodents²³⁶.

2.7.4 Physicochemical properties

Protease function is governed by different factors ranging from their structure to their surrounding chemical environment. These factors seem to be highly depended on the sequence information of a particular protein. Although the proteins under study belong to the same group of enzymes, *plasmodial* proteases and human cathepsins differ from each other. This difference is mainly based on the sequential order of the constituting aa, number and their nature.

Table 2.3: A summary of the different physicochemical properties for FP-2, FP-3 and their homologs

Name	Aromaticity	GRAVY	Instability index	pI	Molecular weight
FP-2	0.13	-0.59	34.98	7.11	55925.19
FP-3	0.13	-0.57	32.76	6.59	56663.97
VP-2	0.13	-0.48	29.78	5.66	55206.20
VP-3	0.13	-0.44	25.96	8.29	56641.75
KP-2	0.13	-0.50	25.13	7.43	56836.52
KP-3	0.13	-0.58	24.76	6.33	55152.13
BP-2	0.13	-0.57	40.40	6.02	54456.60
CP-2	0.13	-0.54	46.76	7.03	54605.15
YP-2	0.13	-0.56	41.37	6.38	55024.47
Cat K	0.10	-0.54	31.48	8.92	23494.57
Cat L	0.12	-0.54	34.58	4.64	24297.87
Cat S	0.12	-0.45	24.29	7.64	23992.04
Mean	0.12	-0.53	32.69	6.84	47691.37
SD	0.01	0.05	7.31	1.17	14352.46

Green = human plasmodial proteases, sky blue = rodent homologs and human cathepsins = yellow.

Despite low sequence identity between human cathepsins and *plasmodial* proteases, the proteins share almost the same aromaticity (0.12 ± 0.01) and GRAVY (-0.53 ± 0.05). Based on the hydrophobic scale by Kyte and Doolittle²³⁷, all the proteins analysed were hydrophilic an indication that they could interact well with water (Table 2.3). All proteins had a positive instability index (32.69 ± 7.31), an indication that they were highly stable. Interestingly, the rodent *plasmodial* homologs had higher values than the rest of the proteases including human

cathepsins. The proteins had a varied range of pI values ranging from acidic to basic (6.84 ± 1.17). In the case of the *plasmodial* proteases, VP-3 had the highest pI value of 8.29 while VP-2 had the lowest (5.66). Overall, Cat L was the most acidic with a pI of 4.64 while Cat K had the highest at 8.92. The observed differential pI profiles might be explained by the localization aspect of the cathepsins. Cat K is predominantly extracellular (osteoclasts) while cat L is the major lysosomal endopeptidase. *Plasmodial* proteases had a higher molecular weight than human cathepsins. This was due to the longer chain length of the mature domains of the *plasmodial* proteases due to the two unique inserts.

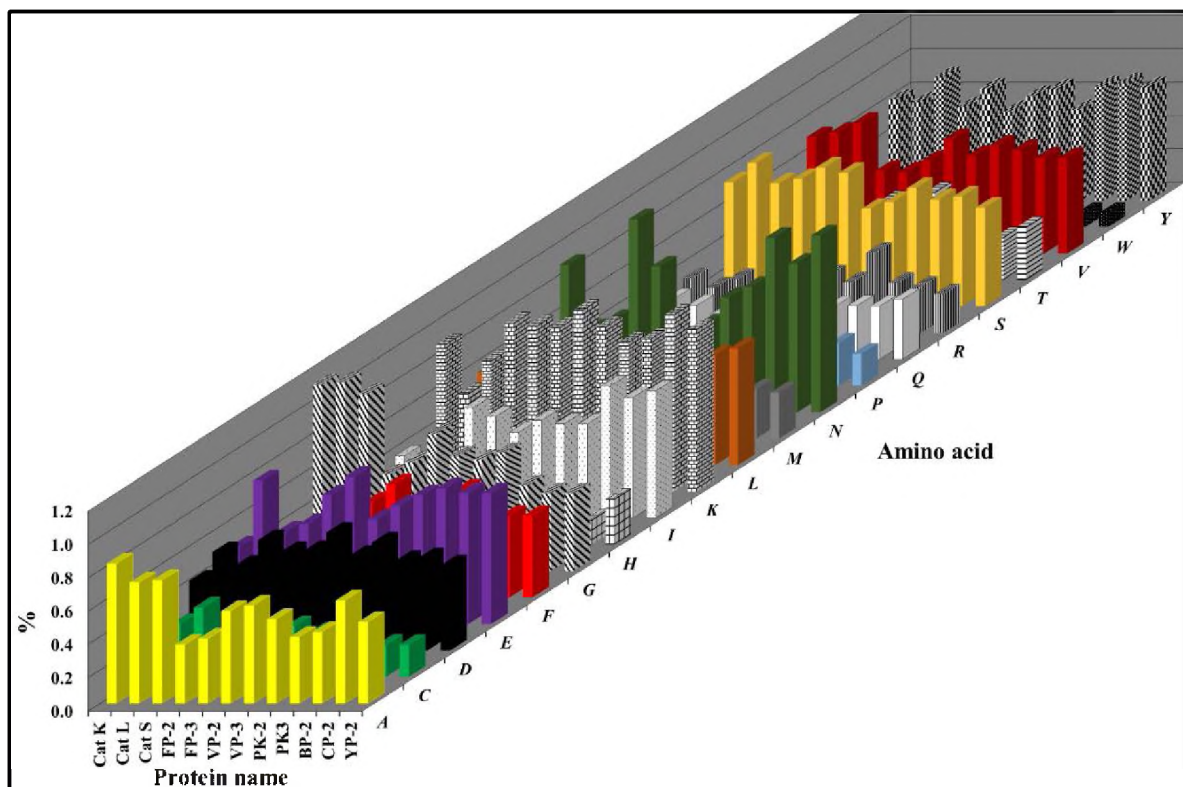


Figure 2.11: Protein composition analysis. The quantitative and qualitative aa composition of FPs and their homologs.

The aa composition varied where the occurrence and distribution of charged aa (Figure 2.11): for positively charged, Lys (K) was more common than Arg (R) in all proteins for negatively charged, the distribution Glu (E) seemed to be equal to that of Asn (N). The occurrence of Gly (G) was most common in human cathepsins than in *plasmodial* proteases. Occurrence of

Phe (F) is lowest in human cathepsins, followed by rodent proteases and the human *plasmodial* proteases. The proportion of Pro is higher in human cathepsins than in the *plasmodial* proteases; Trp (W) was the most infrequent aa which was expected since it is the most bioenergetically expensive aa in nature and is only used when necessary. For polar aa, the occurrence of Ser (S) was higher in all proteins compared to Thr (T). Presently, the correlation between the occurrence of specific aa and individual protein function could not be established and further studies by protein chemists are necessary.

2.8 Chapter conclusion

The current chapter presents an in depth sequence and structural analysis of FPs and their human and *plasmodial* homologs. The database search identified 13 *plasmodial* homologs of FPs were identified from different *plasmodial* species with focus being drawn to the human and laboratory infective species. Four human FPs homologs were identified for comparative purposes. Human Cat H is distinct from other cathepsins, such that this analyses concentrated on Cat K, L and S. MSA identified key features that distinguish *plasmodial* homologs from human cathepsins. Particularly relevant is the distinct nature of the binding pocket subsites leading to separate clustering of the proteases. The effect of this difference for inhibitor selectivity at present remains uncertain. From the physicochemical analysis, no striking differences were observed between the human cathepsins and the *plasmodial* proteases. From the aa qualitative and quantitative analyses, several differences which account to observed individual protein physicochemical properties were noted.

In the next chapter, work involving the interaction of known FPs non-peptide compounds and an identified natural compound from South Africa will aim to determine if these differences can be exploited to attain drug selectivity.

CHAPTER 3

Molecular Dynamics Simulation Studies

Drug discovery is greatly hampered by the limited understanding of the relationship between protein sequence, structure and function. Protein-ligand (substrate or inhibitor) interactions are central to many biological processes. The study of the interactions involved in molecular recognition is important in the modern drug development process. Despite much genomic and structural data, further analysis to define the functions of proteins is necessary. As proteins are structurally dynamic entities, studying their conformational evolutions (structural changes) may provide information of their mechanism for computer aided drug design (CADD). Studying protein dynamics is challenging as most methods provide static information. However, computers can model protein dynamics up to an atomic scale. Here, in silico molecular dynamics (MD) simulations will be used to study FPs and homologs in complex with small non-peptide compounds to provide information on the conformational flexibility of these proteases which can be valuable towards the design of novel antimalarial drugs. The findings will be compared to various structural differences between plasmodial and human homologs identified in Chapter 2.

3.1 Introduction

Life processes are mediated by biomolecular molecules called proteins. These systems consist of a vast quantity of molecules and atoms which are organised into highly defined 3D structures. For proteins to function as biological catalysts, signalling molecules, transporters, sensors and mechanical effectors, they need to be flexible and be able to adapt to the ever-changing conditions within the cellular environment. The activity and interactions of proteins depends on their structure stability, dynamics and flexibility of their constituent atoms, aa, side chains groups, C α -backbone and domains. To understand protein properties and function we need to determine their dynamics at atomic level. Conformation changes may involve domain movements or subtle side chain oscillations and molecular vibrations.

3.1.1 Methods for studying protein dynamics

Two broad approaches to describing protein dynamics include experimental and computational techniques. The chosen approach depends on the type of and complexity of the system. Experimental approaches include electron microscopy (EM)²³⁸, nuclear magnetic resonance (NMR) relaxation experiments²³⁹, X-ray crystallography²⁴⁰⁻²⁴², atomic force microscopy (AFM)^{243,244} and Förster resonance energy transfer (FRET)^{245,246}. However, each can be time consuming and expensive. Also they provide limited resolution in space and time and yield an averaged picture of the protein properties (Figure 3.1).

To describe protein dynamics, computational simulations provide details of proteins at the finest level possible. Biophysics or the “*application of physics laws biologically*” allows three methods of quantifying protein energy, namely quantum mechanics (QM), molecular mechanics (MM) and hybrid models (QM/MM). These are also used in classical MD simulations. Starting at a given point in time they simulate possible changes. MD is the computational study of the positions and velocities of atoms in a system such as a solvated

protein with a ligand due to intermolecular forces by the application of the Newton's laws of motion. The central focus of this chapter will be on MD simulations (at an MM level).

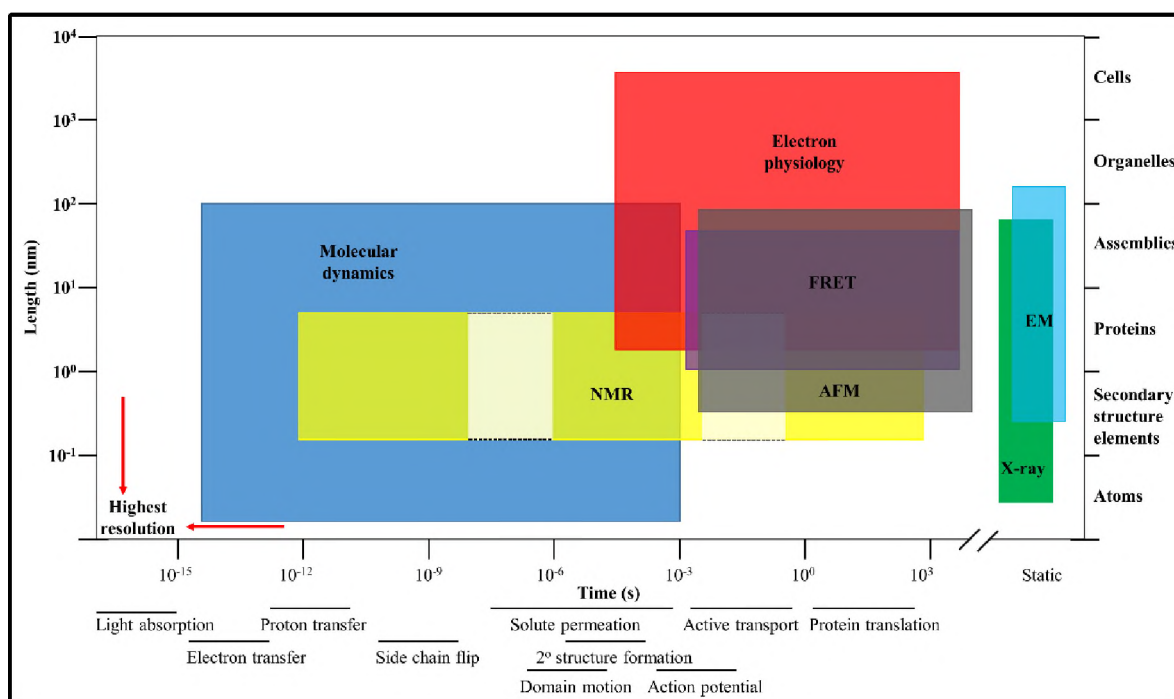


Figure 3.1: Biophysical techniques and their applications. The coloured boxes show the spatiotemporal resolution of each technique. Shown in dotted line are intermediate timescales that NMR is limited leading to inadequate information. Also shown below the x-axis are the timescale of key molecular and physiological processes. Resolution increases in the direction of the red lines. Adapted from Dror *et al.*,²³⁸.

3.1.2 MD limitations

Despite simulating the biophysical behaviour of a system, MD is faced with two major limitations; inadequate sampling due to time limitation thus affecting the precision by which quantities can be estimated; and the lack of appropriate force field functions (where MM is used) and parameters thus affecting the accuracy of the simulations^{239,240}. In relation to time, a good example would be one witnessed during the initial stages of this study. To obtain a single 10 ns run in this study (on a 55.93 kDa protein in a triclinic box of size 17.5 Å containing ~21,000 water molecules) using a modern workstation with 4 processors and a random access memory (RAM) ~32 GB, an average wall time of 192 hours (8 days) was required thus making it impossible for large scale runs to be undertaken. To increase the

sampling effectiveness, several techniques beyond the scope of this thesis have been developed. These include coarse-grained MD²⁴¹, metadynamics²⁴² and conformational sampling²⁴³. However, the implementation of these techniques leads to loss of information compared to the conventional MD processes. Using QM principles, the development of force fields for MM that are accurate for their use in MD simulations is ongoing. It is important to mention that for successful MD run to be realised, the force field chosen must be carefully determined and validated²⁴⁰.

3.1.3 MD simulations history in chemistry and biology

The simulation of many-atom systems predates the advent of modern computers²⁴⁴. Since the 1950s when Alder and Wainwright performed the first proper simulation on the assembly of hard spheres²⁴⁵, MD methods have developed significantly enabling the elucidation of the thermodynamics, structural and dynamic properties of complex biomolecular systems. One of the ground breaking achievements was work done by a South African born scientist involving the computational simulation of protein folding in 1975²⁴⁶. This work formed the foundation of protein MD simulations and led to a sequel of many others that made Michael Levitt and his associates Martin Karplus and Arieh Warshel be awarded with the 2013 Nobel Prize in chemistry. This was due to their contribution in the development of multiscale combined quantum and classical mechanics models for elucidating the course of complex chemical systems through computers.

Since then more complex simulations have been realised, making MD an attractive method in structure-function elucidation. Examples of notable breakthroughs include; thermodynamic fluctuations in a protein (1976)²⁴⁷; the first MD simulation of a protein, bovine pancreatic trypsin inhibitor (BPTI) in 1977 for 9.7 picoseconds (ps)²⁴⁸; dynamics of ligand binding to heme protein (1979)²⁴⁹; a geometric approach to macromolecule-ligand interaction (1982)²⁵⁰; dynamical theory of activated processes in globular proteins (1982)²⁵¹; normal modes and

fluctuations in BPTI (1983)²⁵²; accurate simulations of protein dynamics in solutions (1988)²⁵³; determination of transition paths in macromolecules (1995)²⁵⁴; role of hydration and water structure in biological and colloidal interactions (1996)²⁵⁵; assembly of protein tertiary structures from fragments (1997)²⁵⁶; contact order, transition state placement and refolding rates of single domain proteins (1998)²⁵⁷; unfolding of titin immunoglobulin domains by steered molecular dynamics (1998)²⁵⁸; replica-exchange molecular dynamics method of protein folding (1999)²⁵⁹; energetics of ion conduction through the K⁺ channel (2001)²⁶⁰; Kemp elimination catalysts by computational enzyme design (2008)²⁶¹. Importantly to note is that with the advancement of computational speed and efficiency, MD simulations of more complex systems can be performed at extended time scales of up to millisecond time scale²⁶² (Figure 3.2). Recently, the longest ever MD simulation to be realised was that of a DNA complex done for a duration of 44 μ s²⁶³.

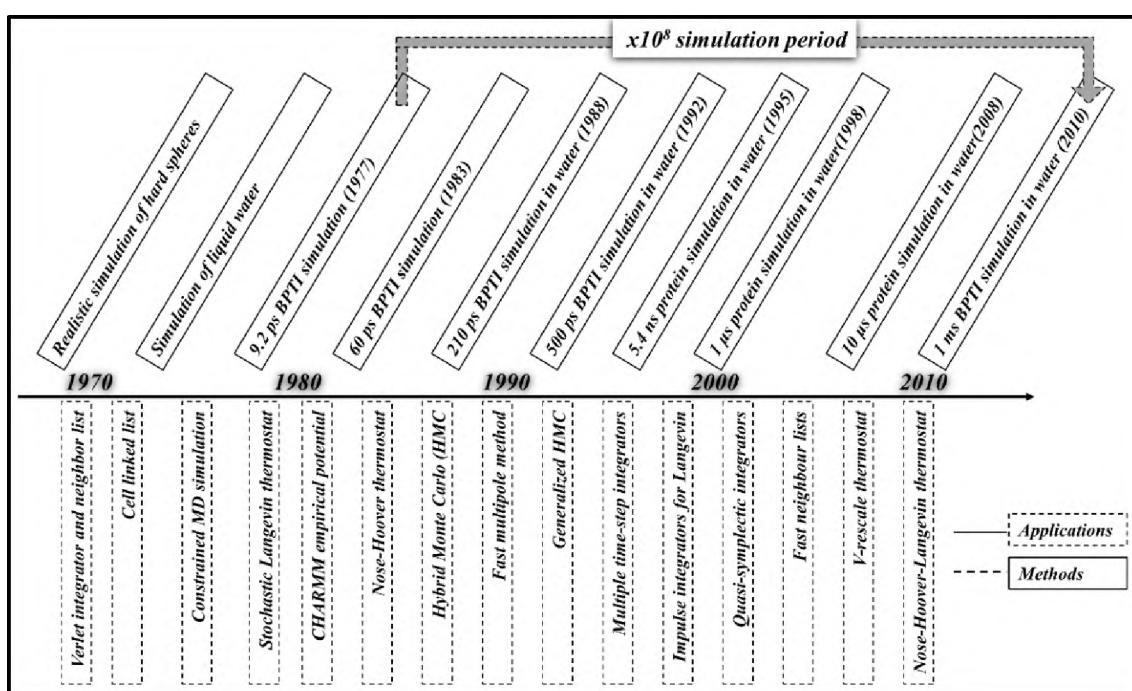


Figure 3.2: The time evolution of key developments in MD simulation and the resulting effects in simulation length of BPTI. The grey dotted arrow highlights the increased capability of simulation period over the last ~30 years. Adapted from work by Bou-Rabee N, 2014²⁶⁴.

Due to their size, MD simulations involving proteins are inherently computer expensive and the resulting accuracy and timescales are dependent on the available resources. Despite the major breakthroughs attained so far, the search of better computational architectures to handle more complex systems and for longer time scales is still ongoing. This include the establishment of parallel computing platforms such as the Generalized-Ensemble Simulation System (GENESIS)²⁶⁵, Titan²⁶⁶, BlueGene/L²⁶⁷ and Anton 2²⁶⁸ supercomputers. In addition, the development of accelerators such as cell processors and graphic processing units (GPUs) has led to the speeding up of non-bonded interaction computations. New MD software with GPUs enabled capability and varied ability depending on the problem at hand has also been developed as shown in Figure 3.3.

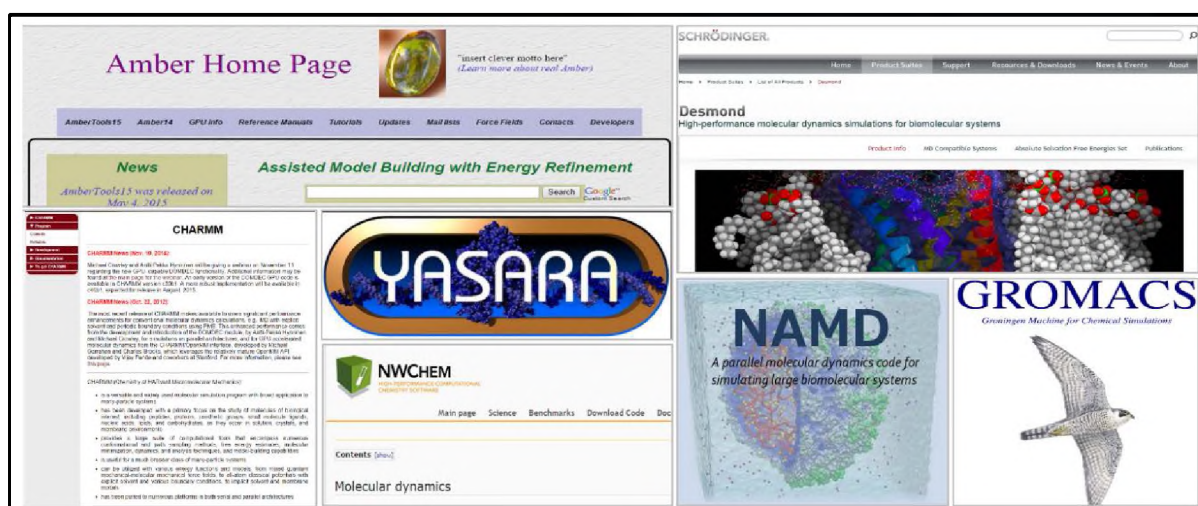


Figure 3.3: Commonly used MD software. A wide range of MD simulation software have been developed. User choice depends on the problem and type of distribution (proprietary or open source).

These MD tools have sophisticated techniques for the sole purpose of attaining reliable MD results. These include the ability to control temperature and pressure. The choice MD software is problem depended and its distribution type (open source under GNU distribution or proprietary). In this thesis, GRoningen MACHine for Chemical Simulations (GROMACS) was chosen as it is a open source software besides its outstanding performance capability, adaptability (can be integrated with other software), state-of-the-art algorithms, and the well

loaded post-dynamic analysis tool kit. YASARA (Yet Another Scientific Artificial Reality Application) is suited for membranous proteins and peptides simulations as it has unique force fields. Despite its user intuitive interface, majority of its modelling and dynamics tools are commercial and cannot be utilised via a command line thus making it impossible to achieve automated workflows.

3.1.4 MD dynamics in drug design

The adoption of MD has changed the study of medicinal chemistry significantly with an emphasis to the elucidation of protein interactions (ligand or protein) and transport of substances within the cell. By definition drug development is an iterative process which has adopted several computational approaches with an aim of lowering the attrition rate (of drugs in screening) by ensuring only potential compounds get through the initial screening stages. Using MD, both qualitative (how, where and when of the drug binding process) and quantitative (strength of interaction and kinetics) information can be obtained. Using the restrictive search words “molecular dynamics” and “inhibitors”, it was determined that more than 14,000 studies have been studied in the last 5 years thus signifying how great MD is embedded in the biomedical research. Initially, the process of ligand binding was conjectured to occur via the lock and key which never accounted for conformational changes. However, this has been replaced by newer models that account for not only definite changes in the structure but random structural adjustments of both the receptor and the ligand²⁶⁹⁻²⁷¹. Examples of areas where MD is appreciated in the drug development process include; 3D (experimental or theoretical model) structure refinement²⁷², determination of protein cryptic and allosteric binding sites²⁷³⁻²⁷⁵, binding mode determination²⁷⁶, the transport process of inside channels, role of mutations in drug resistance or disease pathogenesis²⁷⁷ and the individual changes in aa conformation during the ligand binding process.

3.2 Conventional MD simulations

Computer mediated MD modelling has recently become a very powerful toolbox and an integral part in the study of the dynamics and properties of complex systems. It has found its applications in various fields including material science, chemistry and biology thus demonstrating its versatility. An important application of MD is in drug discovery where insights at atomic scale are deciphered and are used in the R&D of novel therapies. In conventional MD simulations, large systems are handled with classical physics laws (MM, as opposed to QM) thus speeding up the simulations. This is in contrast with the quantum mechanics (*ab initio* and semi-empirical) where the properties of a system is achieved by solving the Schrödinger equation. This latter method although more accurate has a major drawback in that it can only handle systems composed of few atoms at equilibrium. In classical mechanics, all the nuclei in a system are treated as classical particles according to Ehrenfest theorem²⁷⁸. By giving the system an initial set of phase-coordinates and in a step wise manner numerically integrating the laws of motion, a trajectory of the system is achieved. The trajectory consists of time-ordered states of a dynamical system which can be post-analysed to determine the time dependent evolution of the system behaviour. There are several fundamental requirements that have to be met for any MD simulation to be successful as detailed below.

3.2.1 A model of the system

MD simulation studies require the availability of a 3D protein structure that has either been resolved via X-ray crystallography or NMR and in cases where these two are not available, theoretical homology models. Importantly to note is that a careful analysis of the structure is necessary beforehand. Where necessary, the structure is first refined to get rid of local steric clashes using an iterative minimization algorithm²⁷⁹.

3.2.2 Force fields

In molecular modelling, a force field is an empirical potential which replaces the Schrödinger equation used in *ab initio* dynamics. It is mathematical equation consisting of several potential parameters obtained either from semi-empirical QM, *ab initio* or from experimental data and an analytical form describing the interatomic potential energy²⁸⁰. Figure 3.4 illustrates the different types of potentials that constitute a system's potential energy function. Successful definition of classical potentials require two main approximations; the treatment of atom nuclei as having mass and point charges that follow the Newtonian laws of motion. A key reference to reference for the current section is the GROMACS 4.5.5 user manual²⁸¹.

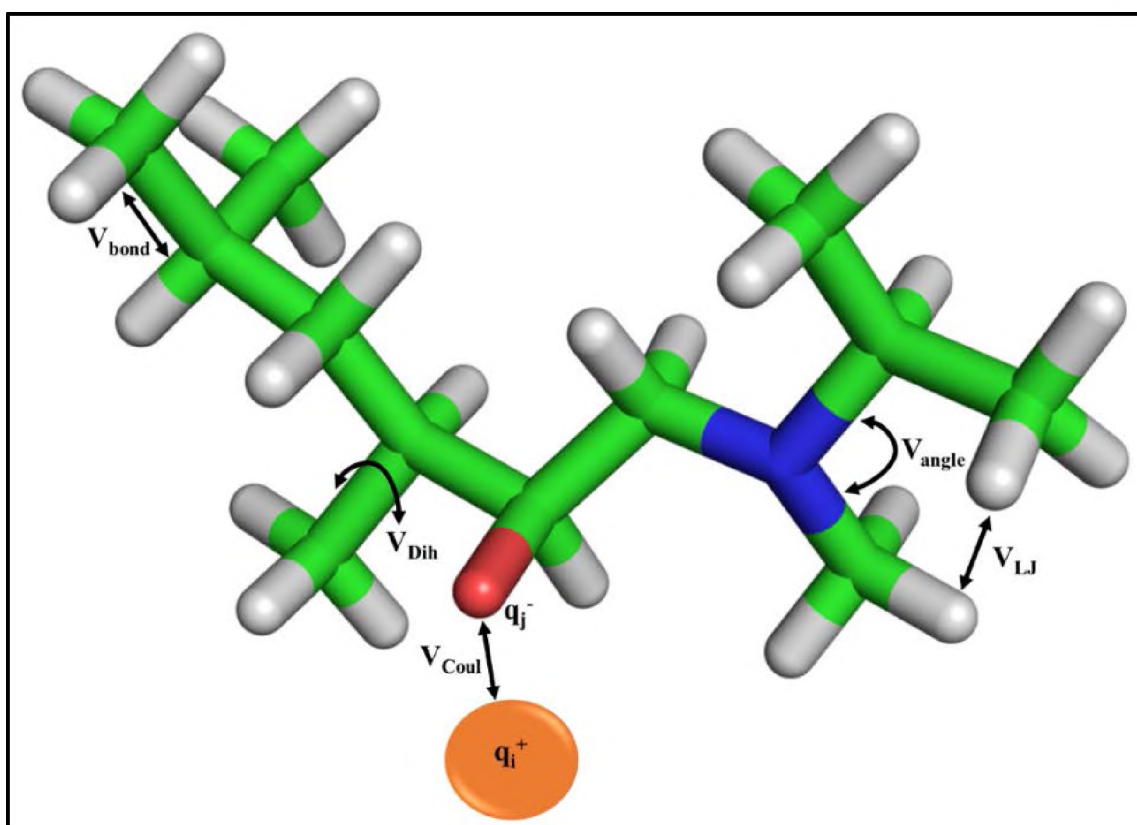


Figure 3.4: A schematic view of force field interactions in a molecular system. The different types of potentials that describe the total energy of a system in a force field. The V_{Coul} (Coulombic potential) describes the electrostatic interactions between oppositely charged points, V_{LJ} (Lennard-Jones potential) represents non-bonded contributions for van der Waals (vdW) interactions. Bonded interaction consist of bond stretching (V_{bond}), angle bending (V_{angle}) and torsion around angles (V_{Dih}).

In *ab initio* molecular dynamics the Born-Oppenheimer approximation is often used which is based on the differential speed between electrons and nuclei thus allowing the decoupling of their motions. The use of QM is often used separately and prior to simulations to investigate force constants and equilibrium distances/angles between atoms of interest. This has aided the generation of additional parameters for various force fields with GRONingen MOlecular Simulation (GROMOS), Assisted Model building with Energy Refinement (AMBER) and Chemistry at HARvard Molecular Mechanics (CHARMM) being the most commonly used force fields in MD simulations involving biological systems. As shown in equation (eqn) 3.1, the sum total of force field terms is made up of different adjustable parameters including an electrostatic term and a functional form of potential energy (potential function) that consists of bonded and non-bonded terms as defined by Lifson and Warshel²⁸².

$$\begin{aligned}
\mathcal{H}(\mathbf{R}) &= V_{bond} + V_{angle} - V_{torsion} + V_{LJ} + V_{Coulomb} \\
&= \sum_{bonds} \frac{k_b}{2} (l - l_{eq})^2 + \sum_{angles} \frac{k_\theta}{2} (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2} (1 + \cos(n\phi - \delta)) + \quad (3.1) \\
&\quad \sum_{pairs(i,j)} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{R_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}} \right)^6 \right]
\end{aligned}$$

where the bonded contributions consists of harmonic functions for bond stretching (V_{bond}) according to Hooke's law, angle bending (V_{angle}) and torsion around dihedral angles (V_{dih}). The non-bonded interactions are from the Lennard-Jones (V_{LJ}) and Coulombic potentials ($V_{Coulomb}$). Denoted with $k_b, k_\theta, \theta_{eq}$ and l_{eq} are bond, angle bond parameters, equilibrium angles and equilibrium bond lengths respectively. The torsion potential is denoted by n , while V_n and δ describes its barrier height and phase. The LJ parameters are denoted by ϵ_{ij} and σ_{ij} ²⁸³. All the MD simulations in this thesis utilised the AMBER force field (eqn 3.2) in its derived form of AMBER96²⁸⁴.

$$\begin{aligned}
E_{total} = & \sum_{bonds} \frac{k_i}{2} (l - l_{i,0})^2 + \sum_{angles} \frac{k'_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2} (1 + \cos(nw - r)) \\
& + \sum_{i=1}^N \sum_{j=i+1}^N \left(4\epsilon_{i,j} \left[\left(\frac{\sigma_{ij}}{R_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)
\end{aligned} \tag{3.2}$$

3.2.3 Integration of Newtonian equation of motion

During MD simulations a series of successive configurations depending on the set time step are generated from an initial configuration whose atoms position and velocities are known. By integrating the laws of motion for each particle, their position and velocities are known in each resulting configuration leading to a trajectory. Hence, by solving Newton's second of motion (eqn 3.3), the time evolution of an atomic system whose individual particles are of mass m_i can be determined along a coordinate r_i when a force F_{r_i} has been applied.

$$\begin{aligned}
F &= ma \\
\frac{d^2 r_i}{dt^2} &= \frac{F_{r_i}}{m_i}
\end{aligned} \tag{3.3}$$

A derivative of the force potential $V(r_i)$ with respect to the coordinate can be equated to the force acting on the particle at each position F_{r_i} .

$$F_{r_i} = \frac{-dV(r_i)}{dr_i} \tag{3.4}$$

By solving equation 3.4 above, the time depended ($t + \delta t$) position and velocities of each particle can be determined and so is the respective forces acting on the particles. To be able to continually integrate the equations of motion during a simulation, suitable algorithms are necessary. Numerous integrators with varied capabilities have been developed. Examples include the Verlet, leap-frog and Velocity Verlet algorithm. These are dependent on the Taylors theorem²⁸⁵ as described below (eqn 3.5) where v (velocity), a (acceleration), b is the first, second and third derivative of the particle position in respect to time (t) in that order.

$$\begin{aligned}
r(t + \delta t) &= r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \dots \\
v(t + \delta t) &= v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \dots \\
a(t + \delta t) &= a(t) + b(t)\delta t + \dots
\end{aligned}
\tag{3.5}$$

To determine new positions ($t + \Delta t$), the Verlet algorithm considers the positions and acceleration of system particles at time t and that of the previous step ($t - \Delta t$). Although it is the simplest of all, it has several drawbacks; precision loss, requirement of an additional input in order to obtain the first updated list of positions, and lack of an explicit term for velocity. An alternative to overcoming these drawbacks is introduction of a velocity explicit expression through the leap-frog algorithm. As the name suggests, particle positions and velocities leap over each other. To calculate new velocities of particles $v(t + 0.5\Delta t)$, it considers the velocity at $t - 0.5t$ and acceleration at time t . The main drawback of this system is that additional calculations are necessary to determine the energy of the system due to the asynchronous nature of individual particle positions and velocities. The development of the velocity-Verlet algorithm has made it possible for the updating of a system position and its velocities without the Verlet and leap-frog associated problems as it calculates new forces from the current position.

3.2.4 Solvation models

Proteins exist predominantly in aqueous environments. Thus, to accurately perform MD simulation involving biological systems, realistic solvent models are necessary²⁸⁶⁻²⁸⁸. The importance of influence of bulk solvent (usually water) on solute molecule has been confirmed by the poor results from *in vacuo* MD modelling. Two broad approaches have been developed; explicit and implicit models²⁸⁹. Implicit models treat the solvent molecules as a uniform polarisable medium having a defined dielectric constant (ϵ) hence they are also known as continuum models²⁹⁰. Several implicit models have been described; surface area and generalised Born^{291,292}. On the other hand, explicit solvent models are the mostly used

and rely on discrete number of solvent molecules around the system being studied. Due to the large numbers of particles and interactions involved the calculations converge slowly compared to those from the implicit model. A large fraction of CPU wall time is dedicated in handling the solvent molecules. Several explicit water models have been defined depending on site points, polarization effects and the type of bonds involved. These include single point charge (SPC), SPC/E (extended), TIP3P, TIPS all which are three-site models. In addition, four-site models such as TIPS2, TIP4P, TIP4P-Ew, TIP4P/Ice, BF, TIP4P/2005 and five-site models TIP5P, TIP5P-E, BNS and ST2 exist²⁹³. In the current work, all runs adopted the explicit approach with the flexible SPC water model.

3.3 Proposed work

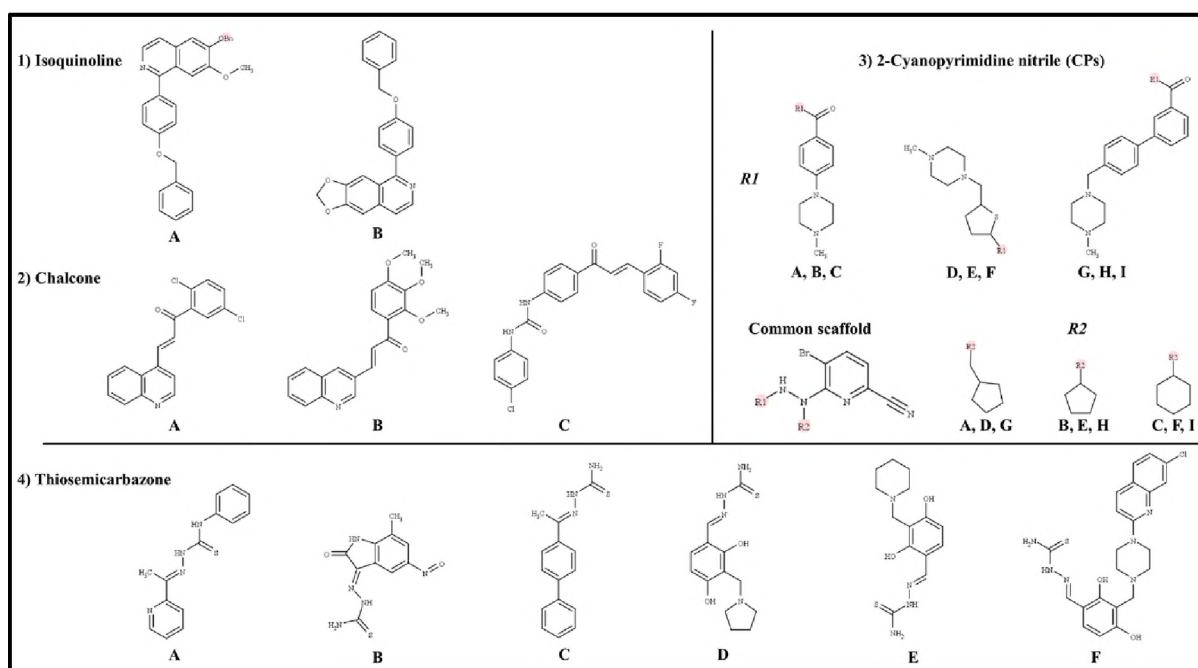


Figure 3.5: 2D structures of known FP-2 and/or FP-3 non-peptidic compounds. From available literature, small compounds with varied inhibitory profiles against FP-2 and or FP-3 were selected for docking. This was to determine their binding mode and evaluate their *in silico* antimalarial activity against FP-2, FP-3 and their homologs from other *plasmodial* species as well as selectivity on human cathepsins. From the docking results, a set of recently reported heteroaryl nitrile derivatives were determined to be the most potent and selected for further studies via MD simulations. Adapted from Musyoka TM *et al.*, 2015²²⁴.

The work in this chapter is a continuation of a project by a preceding MSc student (2013)²¹⁵ in which docking studies using different set of compounds against FP-2, FP-3 and identified homologs (human and *plasmodial*) were performed (Figure 3.5 and Figure 3.6). Initially, a set of known, non-peptide FP-2 and FP-3 inhibitors of the chemical class chalcones²⁹⁴, isoquinolenes²⁹⁵, thiosemicarbazones^{296–298} and 5-substituted-2-cyanopyrimidine nitriles (CPs)²¹⁰ were docked onto the listed proteases. This was to determine their inhibitory potency against the *plasmodial* proteases and selectivity on the human cathepsins. Of these compounds, the CPs were identified as the best inhibitors (Figure 3.5 and Appendix 1H) consistent with available experimental data²¹⁰.

Due to the importance of natural products in drug discovery^{26,299}, a subsequent docking study was conducted on a set of 23 non-peptide natural compounds from SA. A small sterol-like compound 5 α -Pregna-1,20-dien-3-one (5PGA), was identified as a potential hit. The identified hit was used to perform a ligand based virtual screening (LBVS) on the Zinc Is Not Commercial (ZINC) chemical database identifying 186 compounds analogous to 5PGA.

The current investigated the potential of these compounds as antimalarial hits using MD simulations. To reduce the ZINC set of compounds identified via LBVS to only include the compounds with the highest inhibitory potential. Based on the docking energy and the broad activity against the *plasmodial* proteases, a subsequent filtering of the hits resulted to five potential hits with good inhibitory profiles (Figure 3.6 and Appendix 1I and 1J). The main focus of this chapter was to determine the stability between the CPs and the identified hits (5PGA and ZINC) when docked onto FP-2, FP-3 and the various homologs studied previously in Chapter 2 using MD simulations. Also studied was the effect of the various structural and aa composition differences between the *plasmodial* and human cathepsins observed previously (Chapter 2) on the dynamical binding of these ligands. This was to

identify structural and chemical features that could be utilised in improving their activity against the *plasmodial* proteases and or selectivity on the human proteases.

Compound ID	Chemical formula	Mol. Wt.	2D-Structure
SANC00146* (SPGA)	C ₂₁ H ₃₀ O	298.23	
ZINC36371307	C ₃₀ H ₄₈ O	424.27	
ZINC03869631	C ₂₇ H ₄₄ O	384.34	
ZINC04532950	C ₂₆ H ₄₂ O	370.32	
ZINC04579000	C ₂₉ H ₄₆ O	410.35	
ZINC05247724	C ₂₉ H ₄₈ O	412.37	

Figure 3.6: 5 α -Pregna-1,20-dien-3-one and its analogues from the ZINC database. A set of compounds derived from SA natural sources were selected to identify compounds with similar activity as the CPs. Of 23 compounds, one inhibited *plasmodial* proteases with desirable selectivity on the human cathepsins (marked with an asterisk). Through LBVS, up to five potent compounds analogous to the SA natural hit were also identified. This set of compounds together with the CPs (previous figure) were selected for MD simulations. Adapted from Musyoka *et al.*, 2016³⁰⁰.

3.4 Methodology

The stages involved of MD simulation and the tools used are summarized in Figure 3.7. Initially used individually, they were combined into a one standalone tool, *MD_automated.py*, using Python scripting to allow the automation of the whole process of simulation (Appendix 2C). Through available literature, appropriate parameters such as forcefield, box dimension and the various simulation requirements were determined. To optimise the parameters, pre-MD runs were performed and the thermodynamic states of the systems evaluated. The most suitable parameters are specified (Appendix 2C and 2D). For the equilibration and production steps, required system specifications were prepared into single parameter files (Appendix 2D).

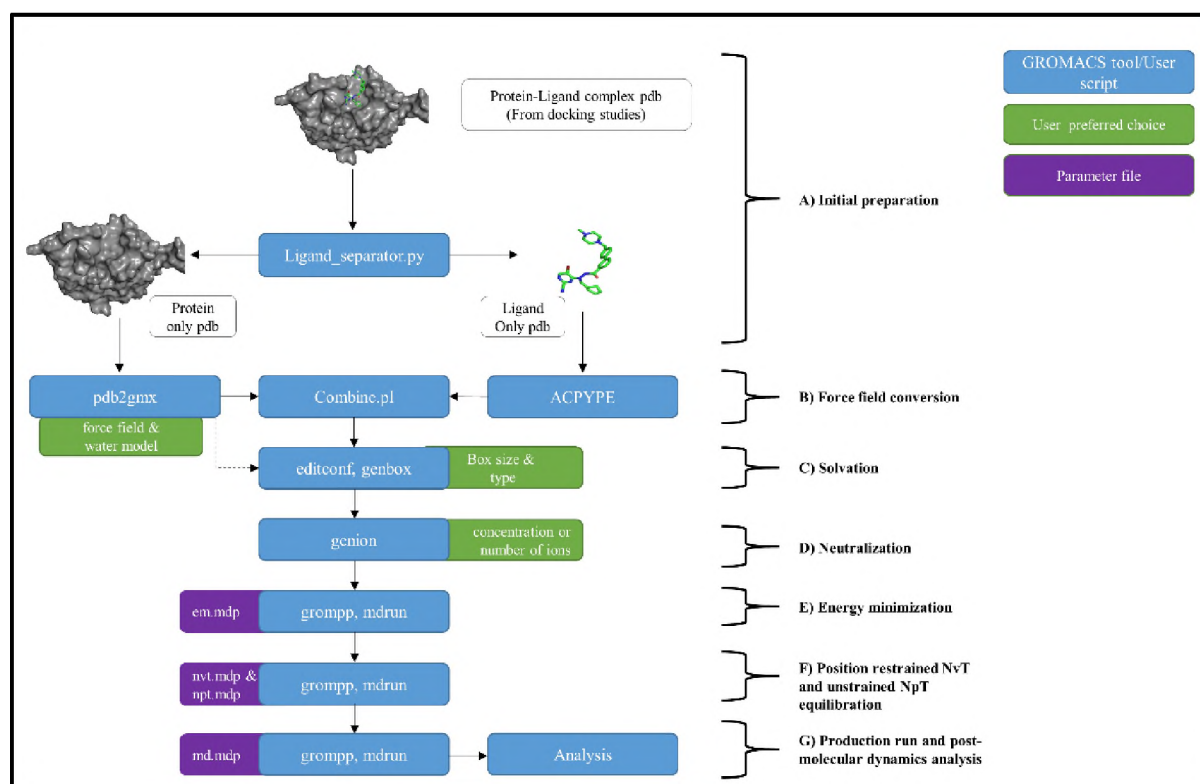


Figure 3.7: MD simulation overview. The different stages utilised for MD simulations. Using Python and Perl programming languages, the different GROMACS tools have been wrapped into an in-house fully automated MD simulation software. Adapted from Brown DK *et al*, 2015³⁰¹.

3.4.1 Preparation of Protein-Ligand Complexes

Using DS-Discovery Studio version 3.5 (Accelrys Software Inc. Discovery Studio Modelling Environment, San Diego: 2011), complexes of the best docking pose for the selected ligands (CPs, 5PGA, and ZINC hits) and the different proteins were prepared. The protein utility module in DS was used to ionize all the protein's titratable residues at a pH of 5.0 and complexes saved in the 3D file format. This was to mimic the lysosomal (human cathepsins) and food vacuole (*plasmodial* proteases) acidic environment where the proteins natively reside²⁰¹. To ascertain that the system was appropriate for MD work a careful evaluation was performed beforehand. This was achieved via visualization using PyMOL. This is to identify the numbering of residues (necessary during analysis step) and where necessary re-number to start from 1“one”. In cases where missing residue(s) are detected, they must be modelled via suitable homology modelling software as GROMACS will not accept incomplete structures. The identification of any hetero atoms (ligands or metallic ions) present is also necessary as these need ligand parametrization.

3.4.2 MD Simulation

3.4.2.1 System set up

A fundamental requirement of MD simulations is the availability of a 3D structure file. As mentioned in Chapter 2, the 3D structures of FP-2, FP-3, Cat K, Cat L and Cat S had already been resolved via X-ray means. The rest of the proteins had their structures calculated via homology modelling and all were used for molecular docking. GROMACS force fields are developed to handle protein atoms, and thus it is necessary to parametrize ligand atoms using external software to a form that can be recognised by the simulation software. Consequently, using a Python script *ligand_separator.py*, each protein-ligand complex prepared by DS was split into corresponding protein and ligand separate coordinate files (Appendix 2E). This was based on the “record name” index (0:6) of all line entries in each protein-ligand complex file where protein entries start with the word “ATOM” while “HETAM” are for the ligand.

3.4.2.2 Preparation of protein and ligand topology files

All the 15 ligands (nine CPs, 5PGA and five ZINC hits) used in this work were not part of the AMBER99 force field parameters. Thus, their force field parameters had to be determined accordingly and placed in a format recognizable by GROMACS. This may be achieved by either using already established external software or through *de novo* build up. A careful consideration is required depending on the complexity of the ligands being studied. *De novo* ligand parametrization is uncommon as it may take years to successfully derive parameters for a single compound. Luckily, several ligand parametrization tools have been established and the choice is dependent on the MD simulation software and force field to be used. These include; Antechamber³⁰² which utilizes the Generalized Amber Force Field (GAFF); CGenFF and The Force Field Toolkit (ffTK)³⁰³ which are for CHARMM compatible parameters; PRODRG³⁰⁴ and automated topology builder (ATB)³⁰⁵ which are for the GROMOS87/96 force field; and Topolbuild and TопоGen specifically for the OPLS-AA (Optimized Potentials for Liquid Simulation – all atom) force field. In this work, the antechamber was selected as the resulting ligand parameters were compatible with the AMBER96 force field which was used for the simulation process. In addition, it has a Python interface, AnteChamber PYthon Parser interfacE (ACPYPE)³⁰⁶ which allows for automated generation of the necessary files. Using a Python script, antechamber and ACPYPE were used to generate the partial charges as well as the force field parameters for each of the 14 selected ligands. At first the ligand atoms were renumbered to correspond to the standard AMBER method. Subsequently, using semi-empirical QM calculations the Mulliken partial charges³⁰⁷ of each ligand atom were calculated. The end result was the generation of a GROMACS compatible (.gro) file, residue topology file (.top) and a corresponding parameter file (.itp). For each protein, a corresponding topology file was generated using the GROMACS utility named *pdb2gmx*. A topology file represents a static description of all atoms and interactions in a system.

3.4.2.3 Explicit solvent simulation parameters

All atom MD simulations were performed using the GROMACS 4.5.5 package³⁰⁸ and employed the AMBER96 force field. This was done in two phases; initially simulations of up to 10 ns were performed for the proteins in complex with the CPs set of compounds and subsequently 20 ns for the 5PGA together with the ZINC analogs. The longer simulation time was necessary for the second set of compounds as 10 ns runs seemed to be inadequate for statistical sampling compared to those of CPs. To create an infinite simulation environment, all MD runs were performed under periodic boundary conditions (PBC) using a triclinic box. Cubic boxes are known to hold more solvent molecules than non-cubic ones of same size dimensions resulting to bulkier system which reduces the computational efficiency of the simulation. To ensure that the box was large enough to accommodate each system, the selection of its dimensions was done via serial trial runs until a dimension of 17.5 Å (L) was determined as the most appropriate for all systems. Allowing adequate space around the protein minimizes periodic artefacts (protein atoms interacting with its neighbours) that will arise from due to unphysical topology during simulation. This allows the creation of an infinite continuous system. Using the flexible SPC water model, an explicit water model was used to solvate the systems. Depending on the overall net charge of each system, a definite numbers of Na⁺ (sodium) and Cl⁻ (chloride) ions were added randomly to the solvent to neutralise the system. Deprived of any constraints, the systems were subsequently subjected to a steep descent energy minimization up to a tolerance of 1,000 kJ mol⁻¹ nm⁻¹. This was necessary in order to remove any steric clashes resulting from the added counter ions and water molecules as well get the systems to a local energy minima. Prior to the production run, systems were equilibrated using the canonical (NVT) followed by the isothermal-isobaric (NPT) ensemble for 200 ps (picoseconds) at each stage. This is necessary so as to scale the systems towards the desired thermodynamic state point. For the NVT ensemble, the systems were slowly heated up to a final constant reference temperature of 300 K within a fixed box

volume ($V=L^3$). Thermostating was achieved by application of the velocity rescaling (V-rescale) algorithm³⁰⁹. For the NPT ensemble, the Parrinello-Rahman barostat algorithm³¹⁰ was used to maintain the pressure of the systems at 1.0 bar in all directions with a pressure coupling constant (τ_P) of 2.0 ps. The values of the isothermal compressibility were set at $4.5 \times 10^{-5} \text{ bar}^{-1}$ for water simulations. The pre-equilibrated systems were then subjected to the production run of either 10 or 20 ns as already explained. Using the leap-frog dynamics integrator, the laws of motion were integrated with a 2 femtoseconds (fs) time step while maintaining temperature and pressure. At each time step, coordinate resetting was performed leading to constraining of solute atoms. All bond lengths during the equilibration and production runs were enforced by applying the LINCS algorithm³¹¹. Long range electrostatic interactions were calculated using the particle-mesh Ewald algorithm³¹² with PBC, a Fourier grid spacing of 0.16 nm and a fourth order cubic interpolation while the cut-off distances for calculation of Coulomb and vdW interactions were set at 1.4 nm. During the sampling process, trajectory snapshots were stored at every 2 ps for structural analysis.

3.4.3 Post-dynamic analysis

Once MD productions runs were completed, the convergence of thermodynamic parameters for each system was monitored for quality assurance. These included temperature, total kinetic and potential energies. Using visual molecular dynamics (VMD) visualization software, the behaviour of each system was examined. To determine the dynamic evolution for the different studied protein-ligand systems over the simulation period, the resulting trajectories were first processed using GROMACS *trjconv* tool. This is necessary to: centre system in the box; remove any periodic artefacts; progressive fitting of system atoms to reference structure (starting 3D coordinate) and to reduce the number of frames. To obtain specific groups (apo, ligand and protein-ligand complex structures) that are important for analysis, corresponding index files were generated. Several GROMACS observables *viz.* root

mean square deviation (RMSD), radius of gyration (Rg) and root mean square fluctuations were calculated from the each trajectory and plotted through Xmgrace of Grace 5.1.21. All the listed analysis steps are included in a Python script, *md_analysis.py* (Appendix 2F). To study the type of interactions between each ligand and the protein binding pocket residues, an *ad hoc* Perl and Python scripts utilizing LigPlot+ subroutines³¹³ was used to analyse the 3D structures obtained during MD runs. For visualization, PyMOL version 1.6.0.0 was utilised.

3.4.4 System specifications

Preparation of protein-ligand complexes was done using a Linux Intel Xenon workstation with an E3-1220V2 quad core processor running at 3.10 GHz, 31.1 GB RAM and Quadro K600/PCIe/SSE2 graphics card. In the case of MD simulations, system set up, solvation, neutralization, and equilibration steps were done on a local cluster. All MD production runs due to their computationally expensive nature were performed on the Tsessebe cluster (Sun) at the Centre of High Performance Computing (CHPC) Unit³¹⁴ in Cape Town, SA.

3.4.5 Drug-likeness of identified hits

Table 3.1: Key physicochemical properties for drug-like molecules

Physicochemical property	Accepted value
Molecular weight (Mwt)	≤ 500 Da
Number of H-bond donors	≤ 5
Number of H-bond acceptors	≤ 10
cLogP	≤ 5
Polar surface area	≤ 140 Å ²
Number of rotatable bonds	≤ 10

Sources: Lipinski et al. (2001)³¹⁵, Veber et al. (2002)³¹⁶, and Keller et al. (2006)³¹⁷.

DrugLito³¹⁸, an open source virtual screening tool was used to determine if the identified hits had drug-like properties. The software determines these properties based on various drug-likeness rules such as the Lipinski's rule of five, Ghose filter, BBB rule, Veber rule, CMC-50 likeness (QED) and the MDDR-like rules (Table 3.1). To determine each descriptor, the software utilises a Java library known as the chemistry development kit (CDK). Prior to

analysis, all the ligand PDB files were converted to a Tripos Mol2 format using the Open Babel software³¹⁹.

3.4.6 MD pipeline

To build a fully automated workflow for MD simulations, GROMACS tools (GROMACS 4.5.5) previously used were incorporated into the JMS (job management system)³⁰¹. This was to achieve a logical flow of all steps necessary for a successful MD simulation in a single pipeline. For sequential flow, stage dependencies were set to ensure that a step could only be executed when its input files have been successfully generated successfully by the antecedent stage. The pipeline consist of up to 18 tools which can be grouped into seven different stages namely 1) initial preparation; 2) force field conversion; 3) solvation; 4) neutralization; 5) energy minimization; 6) equilibration and 7) final production run.

3.4.6.1 Initial preparation

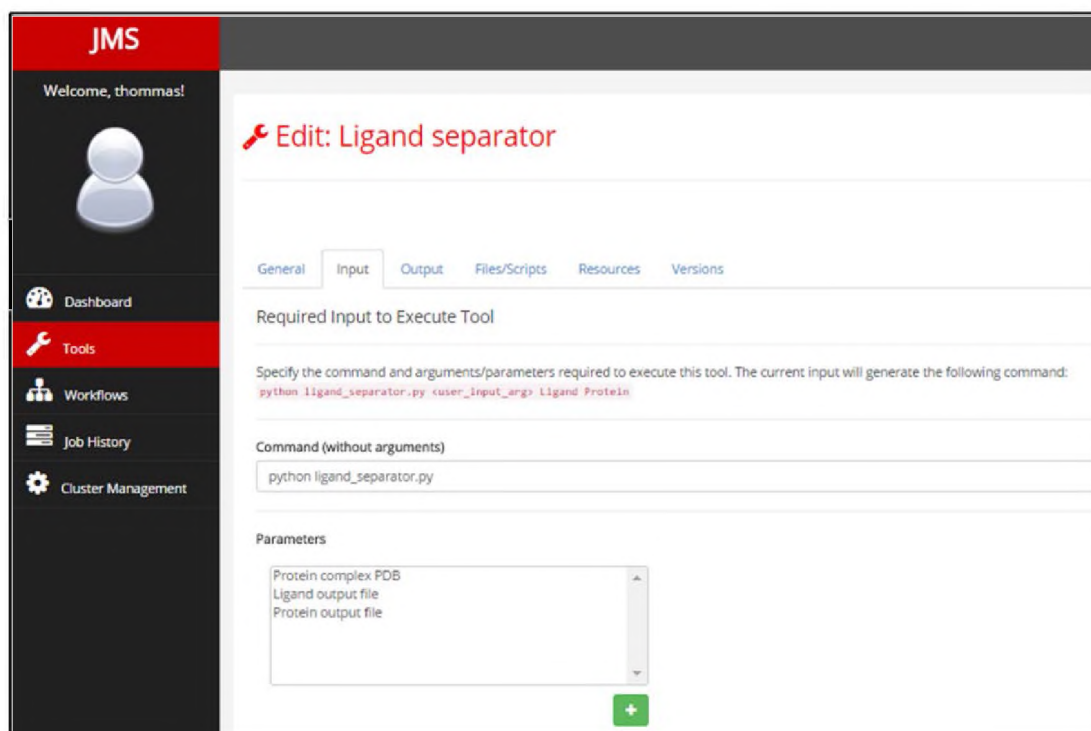


Figure 3.8: The ligand separator interface which acts as the first stage in the MD simulation process. This stage requires the *ligand_separator* Python script.

This is a crucial stage which processes the input file into either one (protein only) or two (protein and ligand) files depending on the nature of problem being studied as defined by the user. No parameters are required to run this particular as everything is handled by a Python script (*ligand_separator.py*) as shown in Figure 3.8.

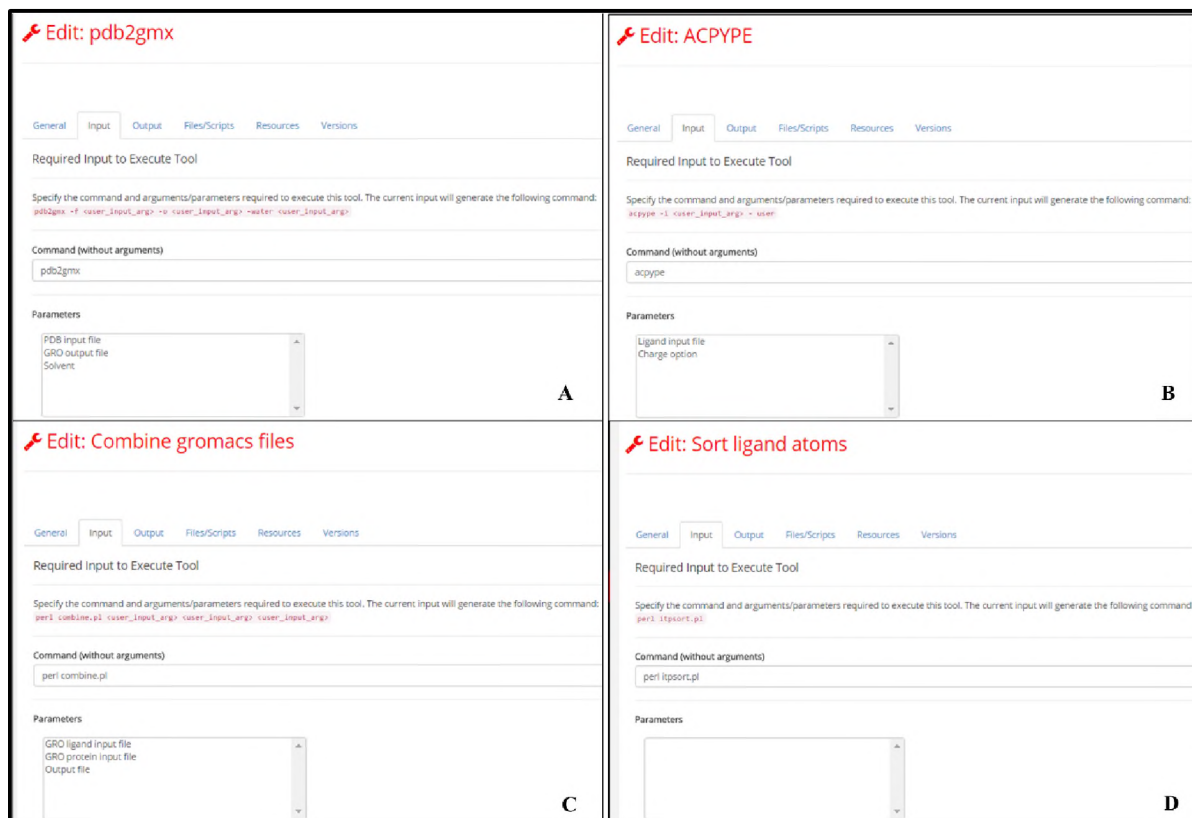


Figure 3.9: The force field conversion interface. **A)** Converts protein from the PDB format to a GROMACS compatible format besides creating topology and parameter files. **B)** *Acypype* tool which only handle any non-protein molecules whose parameters are not included in the GROMACS force field list. **C)** Combines the protein and ligand GROMACS compatible files into a single complex just like the initial PDB input file. **D)** During the conversion by *acypype*, ligand atoms chemical formulae are written in a non-standard form which is unrecognizable by GROMACS.

3.4.6.2 Force field conversion

This processes the input files into GROMACS compatible files which can be handled by the specific force field defined by the user. MD simulations involving protein structures only, require the *pdb2gmx* functionality from GROMACS whereas if a ligand is involved, two additional steps are required. The first involves the *acypype* tool which prepares the ligand file into a format that can be handled by GROMACS via a process called parametrization. A user

must specify the charge option to be used by *acpype* tool otherwise the default AM1-BCC will be adopted. The last step incorporates the combination of the protein and ligand into a single file while taking care of the connection information between individual atoms.

3.4.6.3 Solvation

This is a binary step stage with the first being the creation of a suitable box using *editconf*.

The system requires the user to have prior information regarding the shape and dimension of the box. The other step involves the addition of suitable solvent molecules into the box using *genbox*.

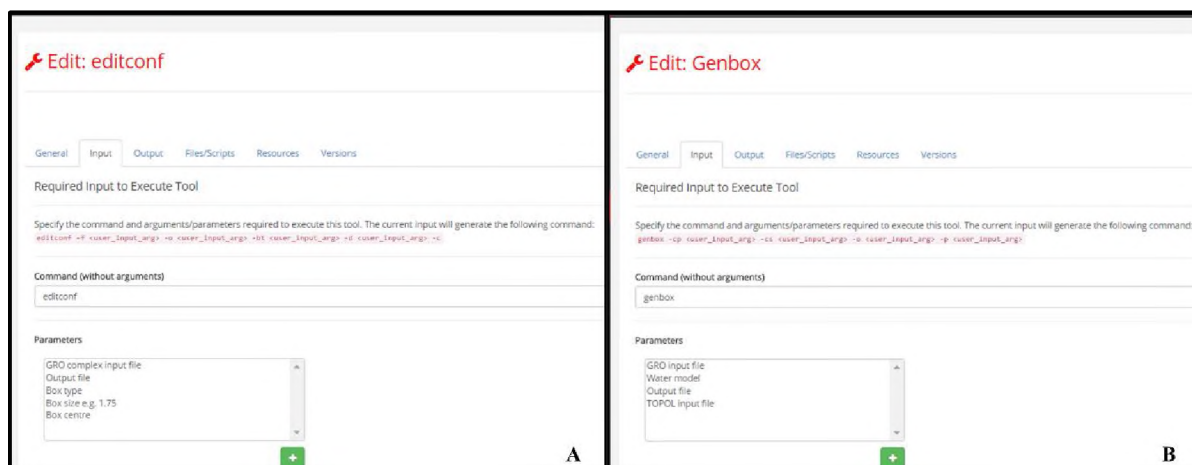


Figure 3.10: The solvation interface where A) the *editconf* tool defines the box shape and its dimensions and B) *genbox* fills the box with solvent molecules as specified by the user.

3.4.6.4 Neutralization

A zero net charge system is a requirement for successful MD simulations. This stage determines the charge of the system and by neutralizes the system by randomly placing enough counter ions in the solvent molecules via *genion*. Only sodium (Na^+) and (Cl^-) ions are allowed. In most cases, a salt concentration of 0.15 M is adequate to achieve the neutralization while maintaining the normal physiological conditions.

3.4.6.5 Energy minimization

A common problem with MD simulations is ‘blowing up’, a situation where simulations fail due to steric clashes resulting from the added water molecules and ions. Also causing this

problem is high energy within the system. The user need to correctly specify the energy tolerance (local minimum) and the minimization method to be adopted (steepest or conjugate descent).

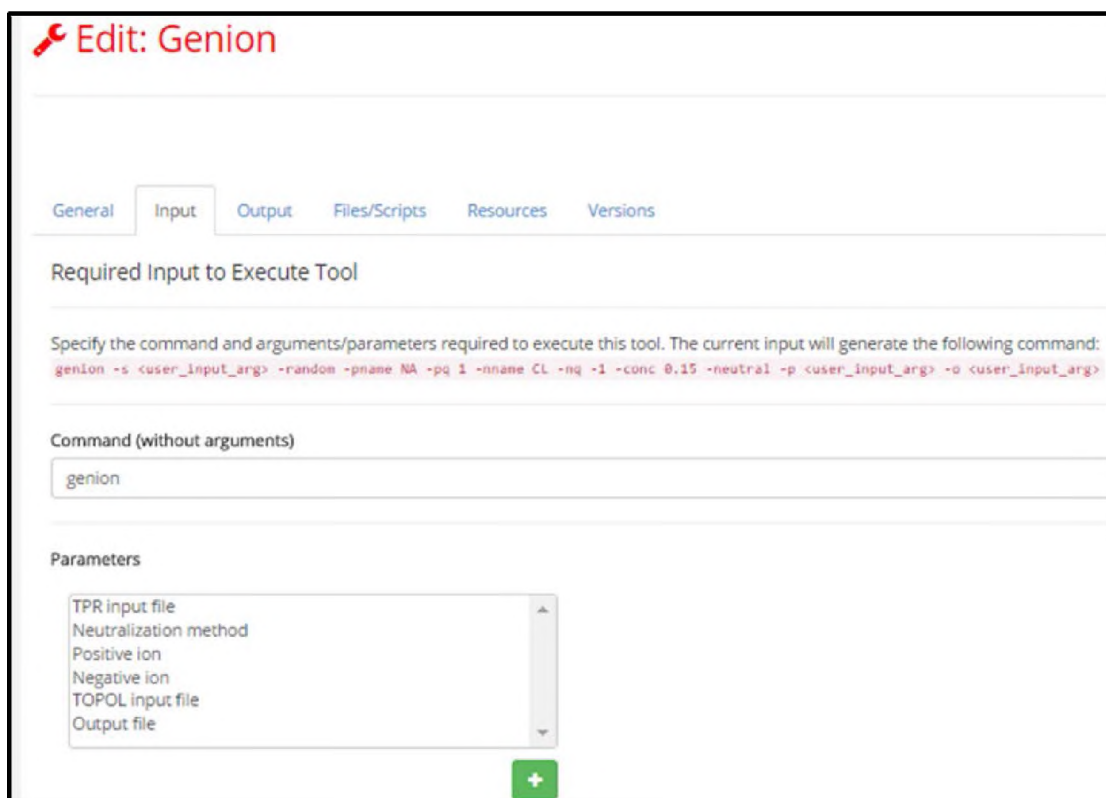


Figure 3.11: *Genion* tool interface and the required parameters for the charge neutralization.

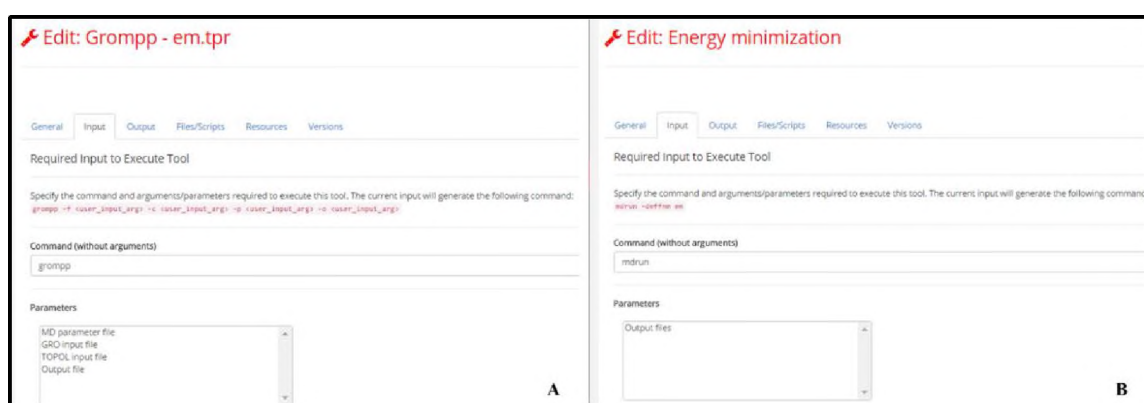


Figure 3.12: The energy minimization tool. **A)** *Grompp* tool requires a user to supply a parameter file with the necessary settings in order to prepare a single input file for the *mdrun* tool which performs the minimization process **(B)**. The type of minimization, energy step and the desired energy tolerance (local minimum) are specified in the parameter file.

3.4.6.6 Equilibration

Prior to production run, the system are equilibrated through a bi-phase process namely *nvt* (isothermal) and *npt* (isobaric) ensemble. Prior to each of this step, *grompp* tool is used to generate a single input file which acts as the input for the *mdrun* tool. The length and temperature among other settings are specified in a parameter file.

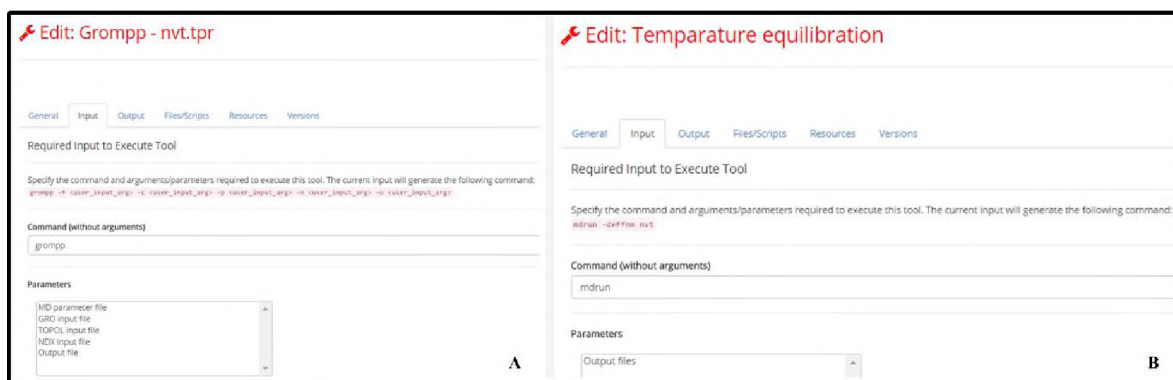


Figure 3.13: Temperature equilibration tool. A) *Grompp* first prepares a single parameter file which acts as an input to the *mdrun* tool for simulation **(B)**.

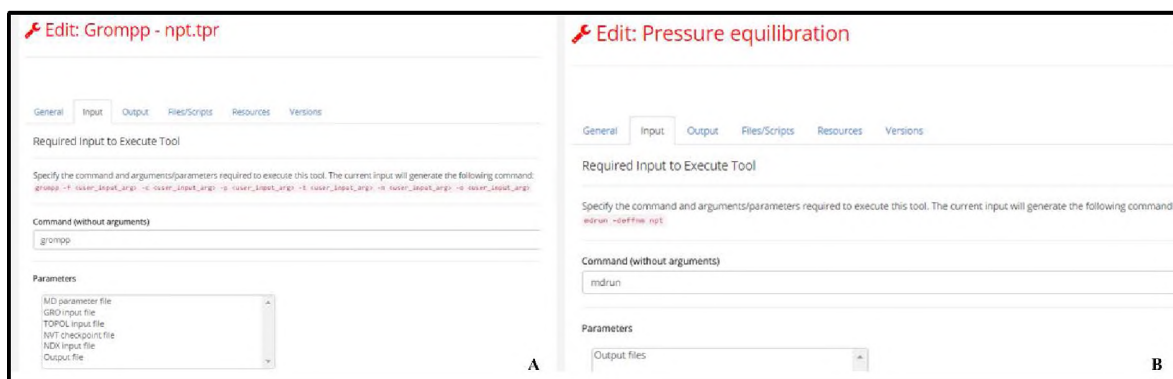


Figure 3.14: The isobaric equilibration tool. A) User defined settings are used by *grompp* to prepare a single input file. **B)** Parameter file is used by *mdrun* to equilibrate the system to the required pressure.

3.4.6.7 Final production run

This is the ultimate stage of MD simulations. Just like any other *mdrun* depended stage, the *grompp* tool is used to prepare a single parameter to be used by the former. All required controls like the extent of the simulation and each time step are supplied in form of a parameter file.

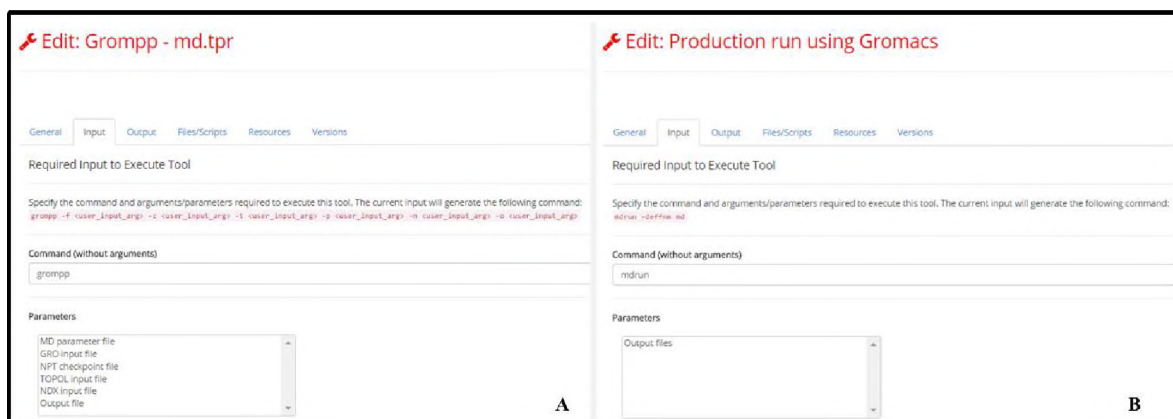


Figure 3.15: Production tool interface. This is used for the actual simulation process and is composed of a *grompp* tool (A) and the *mdrun* simulation tool (B).

3.5 Results and Discussion

In this chapter, nine CPs and one SA natural compound with its five ZINC analogs docked to 12 and 11 proteins respectively were successfully studied using MD simulations. The results herein explore the protein dynamics involved in the binding process to the proteins. Also reported is the successful development of a MD simulation pipeline. The analysis is divided into two sections; first for the CPs followed by that of 5PGA and its selected ZINC analogs. To determine the stability of the protein-ligand complexes during simulation, the convergence of each resulting MD trajectory per system was determined. Using a Python script, the system thermodynamic properties and important observables such as RMSD, RMSF and R_g for all the complexes were determined. The average RMSD of the apo, holo and ligand and the radius of gyration of all systems was determined. The results showed that all the ligands studied were bound to the proteins in a stable manner.

Of great importance was to determine the dynamical footprint of residues involved in the binding of the ligands. All residues forming interactions with the ligands were determined. Also of interest was to establish binding differences of the identified compounds in relation to

the observed structural and sequence differences between the human and *plasmodial* proteases identified in Chapter 2 which can be useful in obtaining drug selectivity.

3.5.1 Quality assurance

As MD simulations are aimed in imitating biological conditions, it is critical to ensure simulations are done in a controlled environment where thermodynamic properties such as pressure, temperature, and density are controlled comparable to the natural environment where the proteins exist.

Table 3.2: The average of different thermodynamic properties during a 10 ns run of different proteins complexed with compound CPG. Simulations were set at a reference temperature of 300 K and a pressure of 1 bar.

System	Temperature (K)	Pressure (bar)
Cat K	299.0 ± 0.2	1.1 ± 0.2
Cat L	300.0 ± 0.4	1.0 ± 0.3
Cat S	301.0 ± 0.3	1.4 ± 0.5
FP-2	300.0 ± 0.3	0.9 ± 0.2
FP-3	299.0 ± 0.4	1.1 ± 0.3
VP-2	300.0 ± 0.5	1.1 ± 0.4
VP-3	300.0 ± 0.4	1.0 ± 0.2
KP-2	300.0 ± 0.4	1.0 ± 0.3
KP-3	300.0 ± 0.4	1.2 ± 0.1
BP-2	300.0 ± 0.4	1.0 ± 0.3
CP-2	300.0 ± 0.4	1.1 ± 0.2
YP-2	300.0 ± 0.4	1.2 ± 0.2

For all simulations performed herein, the systems behaviour and properties were analysed to ensure the reliability and integrity of each system were maintained throughout the simulations. Table 3.2 shows the average of the main properties of the different proteins in complex with CPG as determined over a 10 ns simulation. The convergence of the kinetic energy profiles of FP-2, FP-3 and the human cathepsins when in complex with compound CPG, CPH and CPI is shown in Figure 3.16. The results indicated that the simulations proceeded in the desirable thermodynamic conditions and further analysis could be performed.

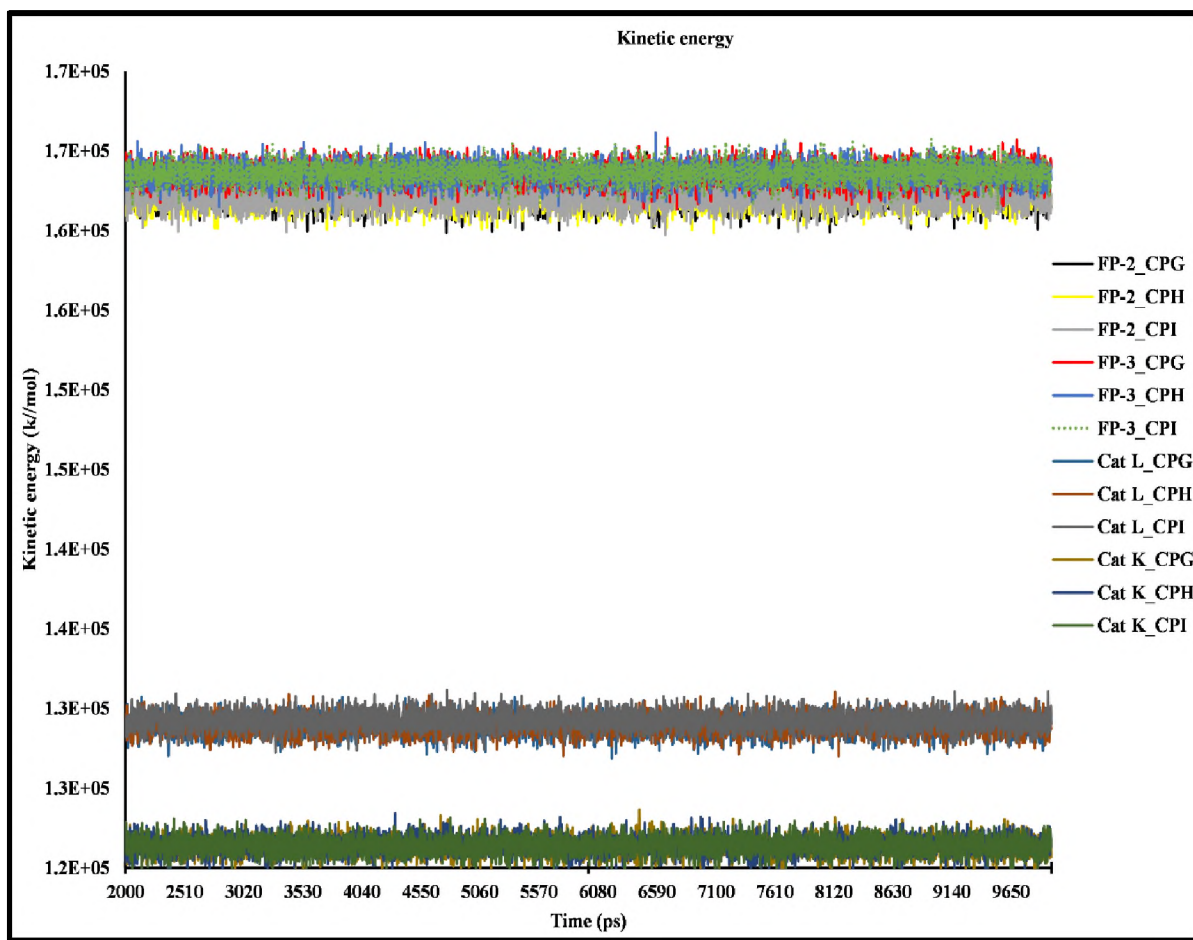


Figure 3.16: Kinetic energy of FP-2, FP-3 and human cathepsins when complexed with compound CPG, CPH and CPI during the last 8 ns of MD simulation.

3.5.2 Visualization

Using VMD, the trajectories of each simulation were visualised to establish the stability of the ligand within the binding pockets of the proteases. Despite minimal flip-flop movements at the rotational bonds, all the ligands remained stably bound to the interacting aa of each protein. Figure 3.17 is an example of a single frame of a simulation system of FP-3 in complex with CPG.

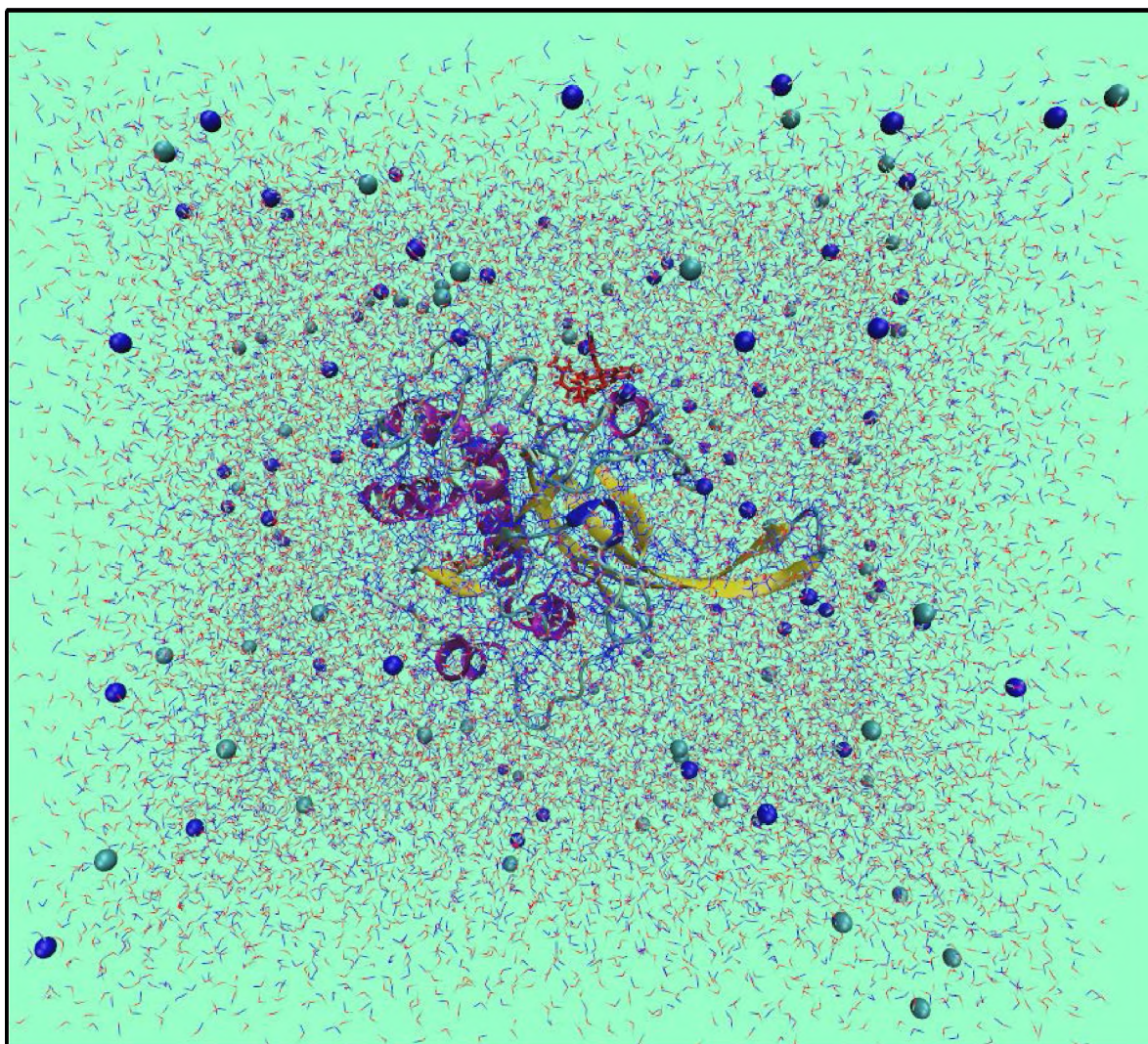


Figure 3.17: A triclinic simulation box of FP-3 (cartoon) and CPG (red) submerged in a box of water molecules (short lines) with Na⁺ (light green) and Cl⁻ (blue) ions.

3.5.3 CPs

Through a sequential lead optimization process on a group of compounds belonging to the heteroaryl nitrile class, Coteron *et al.*, identified a series of derivatives herein abbreviated as CPs (5-substituted-2-cyanopyrimidine nitriles). These compounds have been vouched as the most potent hitherto. Wet laboratory assays showed that they had picomolar to nanomolar inhibitory activity against FP-2 and FP3 and whole *plasmodium* parasites respectively²¹⁰. From docking studies three compounds namely CPG, CPH and CPI had the lowest binding energies.

3.5.3.1 RMSD

In MD simulations, RMSD is the most commonly used quantitative measure of the global conformational diversity in reference to a starting structure using the C α atomic co-ordinates allowing for the determination of spatial differences in an ensemble of structures. This is by solving the following equation (eqn 3.6);

$$RMSD(t) = \sqrt{\frac{1}{M} \sum_{i=1}^n m_i |r_i(t) - r_i^{ref}|^2} \quad (3.6)$$

where $M = \sum_i m_i$ and $r_i(t)$ represent the position of an atom i at the time t after least square fitting the structure to the reference structure (r^{ref}). The RMSD of the different systems increased rapidly for the first 2 ns of each run and stabilised thereafter. Occasionally, minimal fluctuations were observed and were linked to the characteristic loop regions present in all the proteins under study (Figure 3.18a,b). For the apo structures; human cathepsins exhibited the lowest RMSD values ranging from 0.10 to 0.14 nm compared to the *plasmodial* homologs which had RMSD values ranging from 0.14 to 0.24 nm. Similar results with FP-3 have been obtained previously²⁰⁹. Interestingly, rodent *plasmodial* proteases exhibited a higher RMSD values than human *plasmodial* homologs in most cases. This may indicate that the former may have greater fluctuations around the loop regions compared to the latter (discussed under RMSF). It has not yet been established if the observed higher RMSD values associated with the rodent *plasmodium* homologs are in any way associated with the corresponding higher instability index values observed previously (Chapter 2). To determine the effect of ligand binding onto the different proteins, the RMSD of the holo forms of each corresponding protein was determined. From the holo RMSD values, the cathepsins had values ranging from 0.15 to 0.19 nm while the *plasmodial* homologs had values ranging from 0.18 to 0.29 nm. This slight change in RMSD between the apo and holo forms may be linked to the small size

of the CPs. Shown below in Figure 3.19 are the trajectory plots of different apo systems during the last 6 ns of MD simulation.

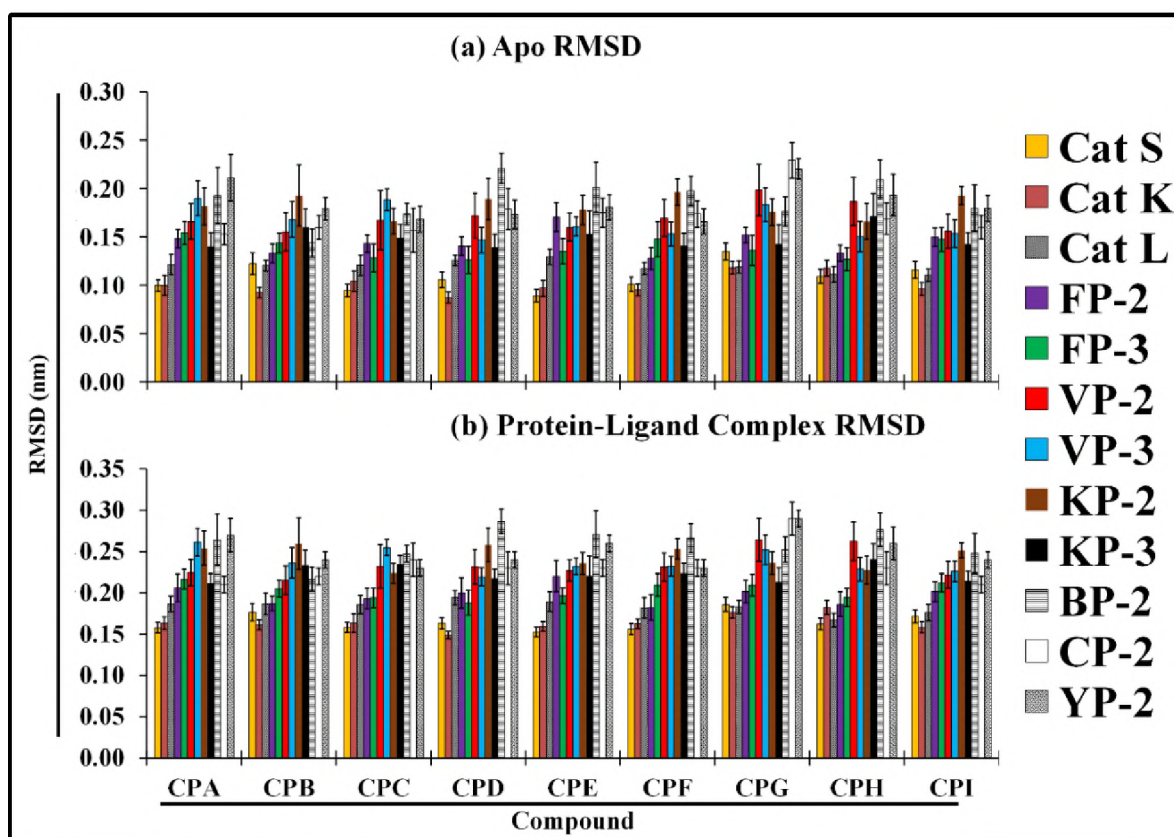


Figure 3.18: Global conformational diversity of *plasmodial* and human proteases when in complex with CPs. The average and SD of the RMSD of the apo (a) and holo (b) systems as determined by the `g_rms` tool for the simulation period from 4 – 10 ns. The error bars indicate the standard deviation of RMSD per system over the last 6 ns of simulation. Adapted from Musyoka, TM *et al.*, 2015²²⁴.

The different ligands exhibited different RMSD values with the largest being from compounds CPG-CPI due to their high number of rotational bonds (8 or 9) compared to the rest which had either 6 or 7 (Figure 3.20). Although these ligands had high RMSD values, they have minimal influence on the overall complex RMSD as seen in Fig. 3.18. The process of ligand binding onto a protein may lead to its stabilization or otherwise and from the small difference between the apo and holo RMSD values these ligands never destabilised the proteins.

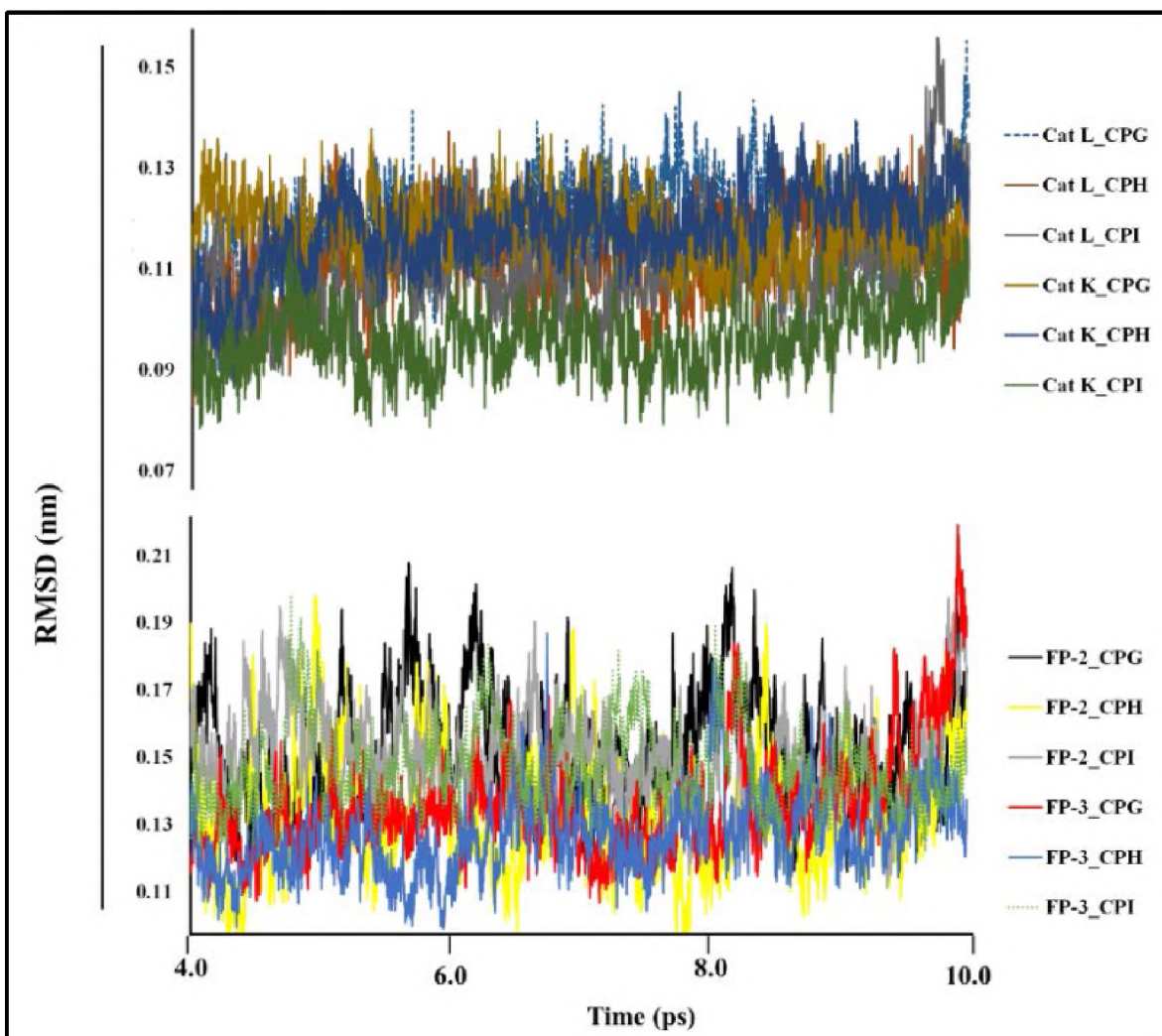


Figure 3.19: Trajectory plots showing RMSD fluctuations of the apo structures of cathepsins (Cat K and L) and falcipains (FP-2 and 3) during the last 6 ns of a MD simulation.

3.5.3.2 R_g

Using GROMACS *gmx_gyrate* tool, the R_g of each molecule about the x, y and z axes as a function of time was calculated by solving the following eqn. 3.7.

$$R_g = \left(\frac{\sum_i |r_i|^2 m_i}{\sum_i m_i} \right) \quad (3.7)$$

where m_i is the mass of atom i and its position in respect to the center of mass is shown by r_i .

This helps us to determine the intermolecular compactness or spread of a molecule as stable structures show a steady R_g and vice versa³²⁰. From Figure 3.21, all proteins were highly

stable with no unfolding witnessed throughout the simulation. Human cathepsins had the least Rg with values of 1.68 ± 0.3 nm. The *plasmodial* homologs had nearly identical Rg values of 1.83 ± 0.2 nm. The difference in Rg between the human cathepsins and the *plasmodial* homologs is as a result of the presence of the arm region only present in the latter.

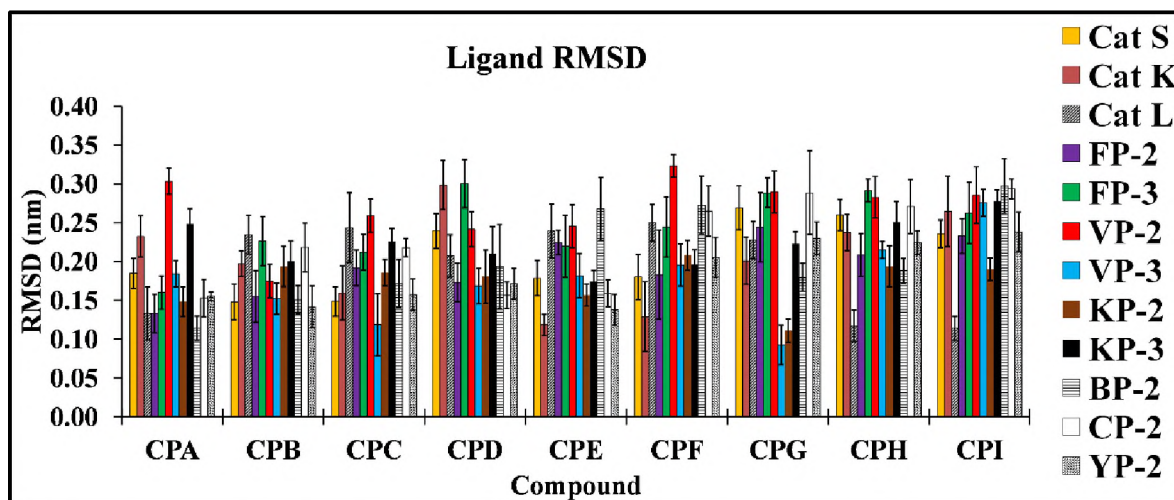


Figure 3.20: Ligand RMSD fluctuations. The average RMSD of the different ligand for the last 6 ns of MD simulations. The error bars indicate the standard deviation of RMSD per system over the last 6 ns of simulation. Adapted from Musyoka, TM *et al.*, 2015²²⁴.

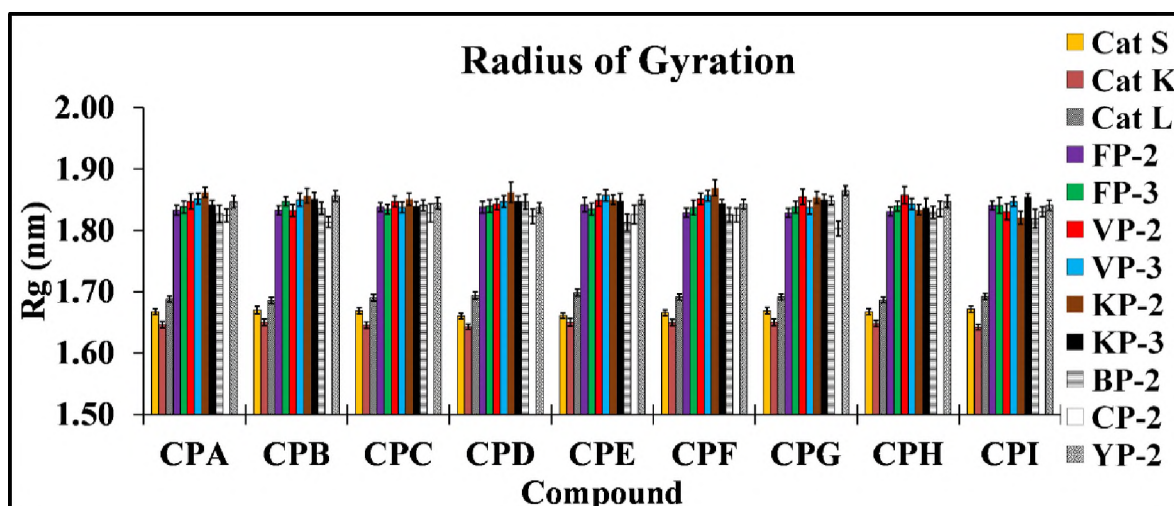


Figure 3.21: Compactness of the different protein-ligand complexes. The radius of gyration of the different systems during the last 6 ns of MD simulations. Error bars indicate Rg standard deviation per system. Adapted from Musyoka, TM *et al.*, 2015²²⁴.

3.5.3.3 RMSF

Using the *g_rmsf* tool in GROMACS, the local fluctuation of all aa during the MD simulations were calculated.

This is achieved by solving eqn. 3.8;

$$RMSFi = \sqrt{\frac{1}{T} \sum_{t_j=1}^T |r_i(t_j) - r_i^{ref}|^2} \quad (3.8)$$

where T is the averaging time while the reference position of particle i is denoted by r_i^{ref} . From the RMSF plots, all aa within loop regions exhibited huge local conformational fluctuations with the *plasmodial* β -hairpin (aa ~175 - 200) residues (Chapter 2) exhibiting the largest fluctuations. As the human cathepsins lack this characteristic feature, they had lower fluctuations around this area. The rodent *plasmodial* homologs exhibited larger MD fluctuations within the loop regions a fact that may explain the reason why they showed higher RMSD values than the human *plasmodial* forms (Figure 3.22).

All residues that form the binding pockets (refer to Chapter 2) were highly stable (< 0.1 nm) an indication that the recorded local fluctuation movements could not interfere with compound binding process (Figure 3.23).

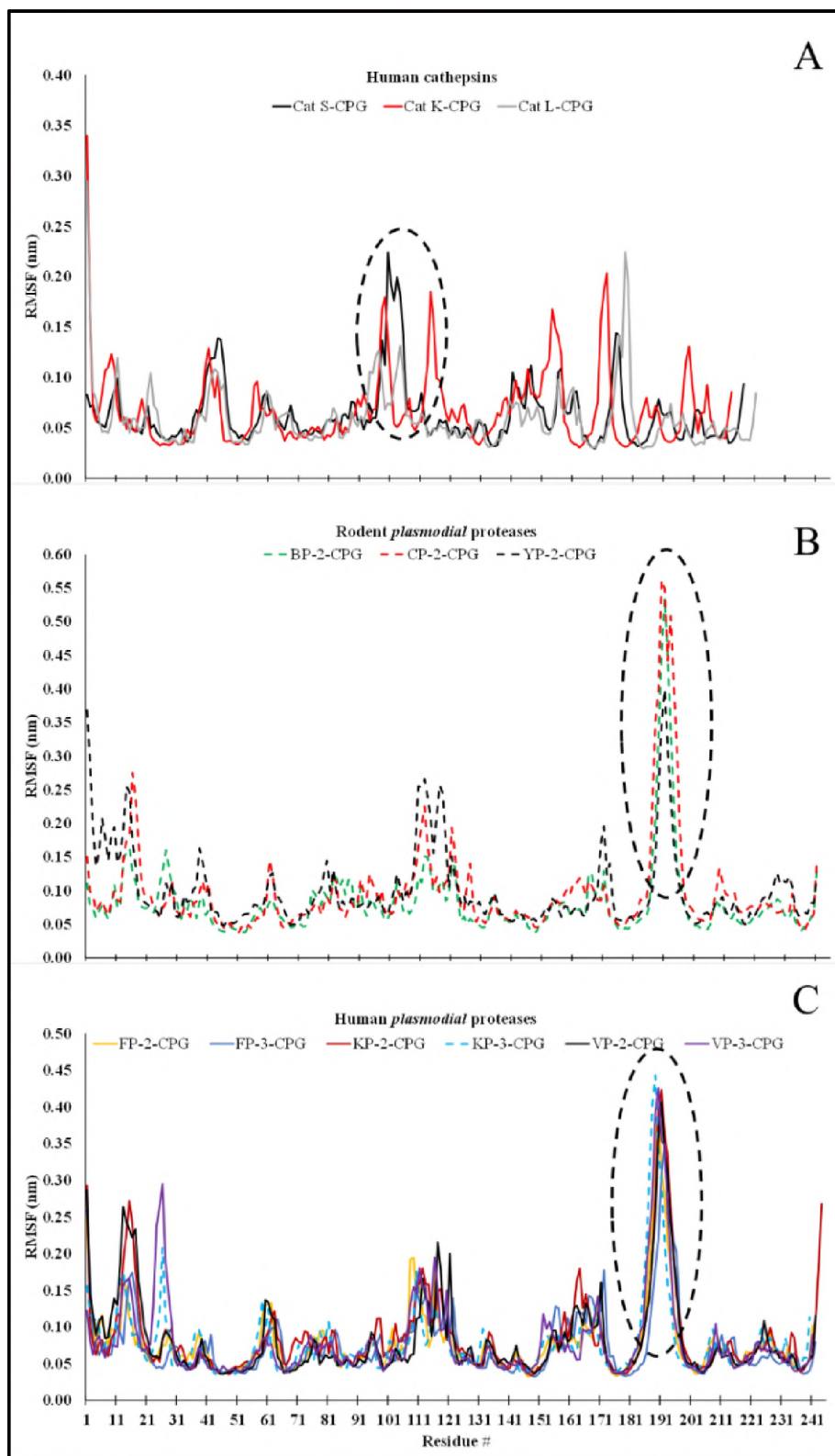


Figure 3.22: Local residue fluctuations of A) cathepsins, B) rodent and C) human *plasmoidal* holo forms (complexed with CPG). RMSF plots depicting the average individual residue changes during the last 6 ns of a MD simulations. Highlighted with broken circle is the residues forming the β -hairpin.

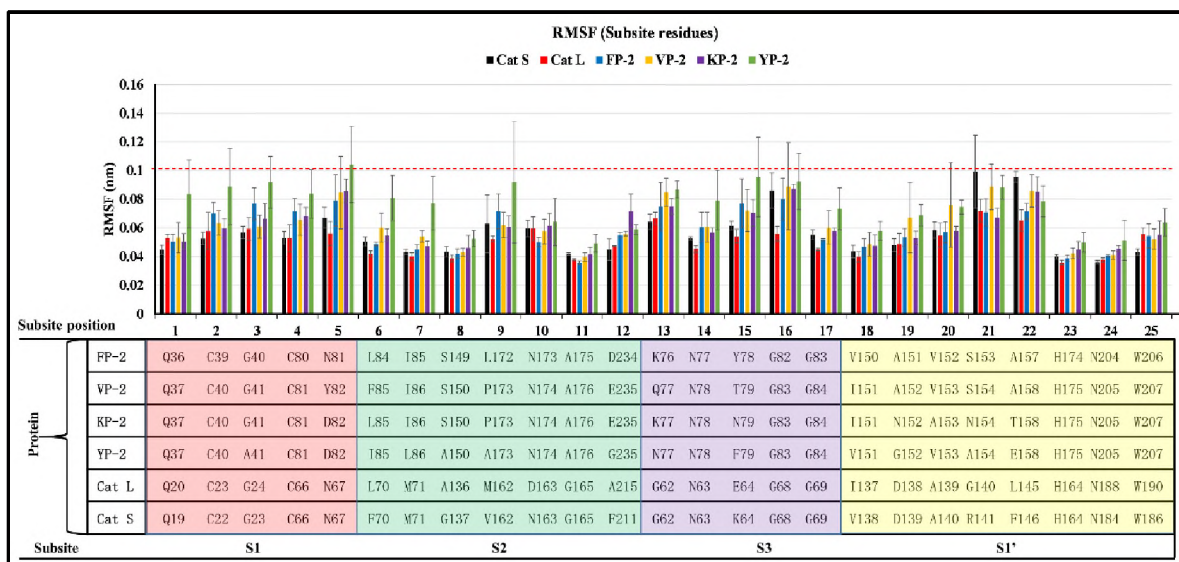


Figure 3.23: The average local fluctuations of the subsite residues for a subset of proteins (shown on the table) when bound to CPG, CPH and CPI. The red line depicts the least average global fluctuations (RMSD).

3.5.3.4 Solvent Accessible Surface Area (SASA)

To determine structural packing, the GROMACS tool `g_sas` was used to determine the hydrophobic, hydrophilic, total and SASA. The SASA factor is used to determine protein stability³²¹ with greater thermodynamic stability associated with lower SASA values.

Table 3.3: `g_sas` output of the different systems

Protein	Hydrophobic nm ²	Hydrophilic nm ²	Total SASA nm ²	D Gsolv kJ/mol/ nm ²
Cat K	39.87 ± 1.19	73.85 ± 1.41	113.71 ± 2.06	340.33 ± 12.81
Cat L	42.45 ± 0.86	72.71 ± 1.31	115.16 ± 1.67	348.46 ± 7.94
Cat S	40.32 ± 1.02	72.52 ± 1.35	112.84 ± 1.84	351.72 ± 8.70
FP-2	48.23 ± 0.94	82.53 ± 1.53	130.76 ± 2.01	413.85 ± 9.31
FP-3	49.52 ± 1.39	83.07 ± 1.62	132.60 ± 2.68	428.69 ± 12.98
VP-2	50.97 ± 1.39	82.25 ± 1.62	133.22 ± 2.68	428.69 ± 12.98
VP-3	52.37 ± 0.98	83.83 ± 1.73	136.20 ± 2.20	444.05 ± 9.99
KP-2	52.20 ± 1.02	85.60 ± 1.28	137.81 ± 1.75	451.85 ± 8.66
KP-3	49.45 ± 1.02	86.56 ± 1.38	136.01 ± 1.85	428.36 ± 9.29
BP-2	44.42 ± 1.34	87.66 ± 1.24	132.08 ± 2.11	414.97 ± 12.52
CP-2	47.74 ± 1.24	84.01 ± 1.26	131.75 ± 1.91	414.72 ± 10.64
YP-2	49.05 ± 1.57	89.99 ± 1.54	139.04 ± 2.33	460.17 ± 12.56

From Table 3.3, human cathepsins exhibited lower hydrophobic, hydrophilic and SASA values compared to the *plasmodial* proteases thus an indication they are more thermodynamically stable compared to the latter. This was in agreement with the results obtained from the Rg, RMSF and RMSD analysis.

3.5.3.5 Binding mode

The binding process of a ligand onto a protein binding site is determined by an array of intermolecular interactions between the aa lining the binding pocket and the ligand atoms. In addition, the stability of a bound ligand on the binding pocket is controlled by interplay of specific and non-specific forces. These forces are depended on the nature of chemical groups within a ligand and those of the surrounding environment. The size of the binding pocket and subsites also determines the entry of the ligand. To gain more insights on the observed activity, the binding modes of compound CPG, CPH and CPI (compounds with lowest docking energies = best binders) were evaluated. From structure visualization using PyMOL, the ligands fitted well onto the S1, S2 S3 and S1' subsites (Figure 3.24). Their extended nature enabled them to interact with virtually all subsite residues within the trench-like binding pocket. A detailed analysis on the effect of these interactions to the binding energetics will be discussed at length in the following chapter (Chapter 4).

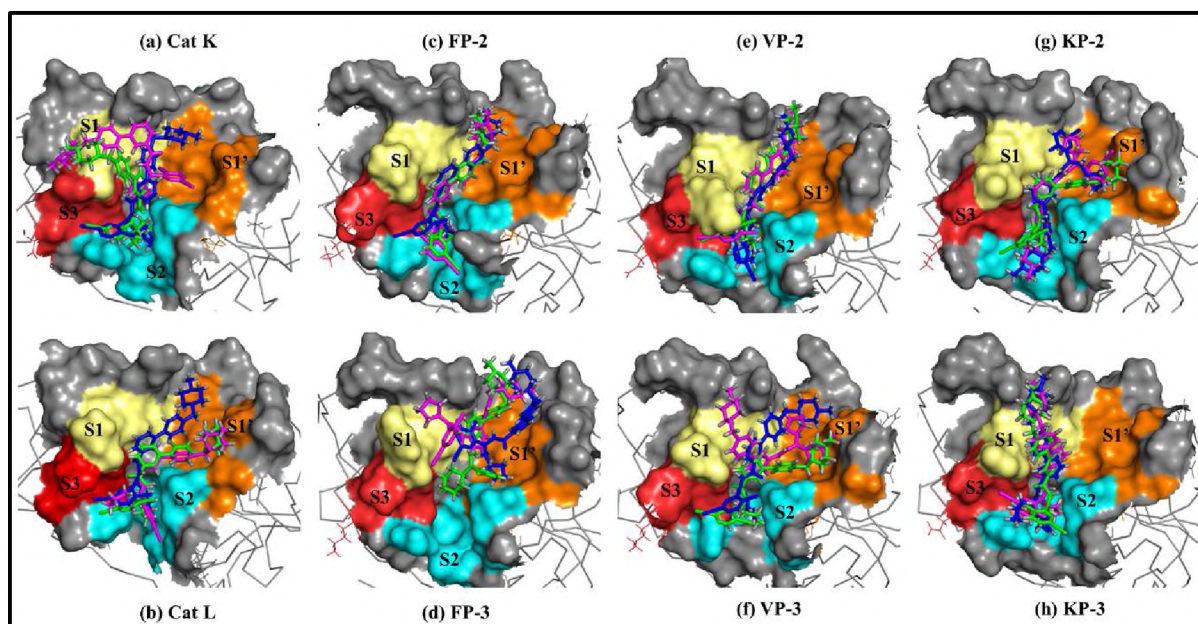


Figure 3.24: The docking pose of CPG (blue), CPH (magenta) and CPI (green) in the binding pocket of human cathepsins and *plasmodial* proteases. Used in Musyoka, TM *et al.*, 2015²²⁴.

3.5.3.6 Structural chemical features of binding

One goal of MD simulations is to determine if the interactions that are critical in the binding of a ligand are stable throughout the simulation period. Here in, the interactions of CPs with protein residues were determined at time 0 ns (docking state) and thereafter during the simulation at a 2 ns interval. Using a Perl script utilizing LigPlot++ subroutines, the specific aa-ligand atoms interaction fingerprint were identified throughout the simulation (Figure 3.25). From the interaction fingerprint (Table 3.3, Appendix 1K), the major interactions involved in binding of the ligands were vdW forces.

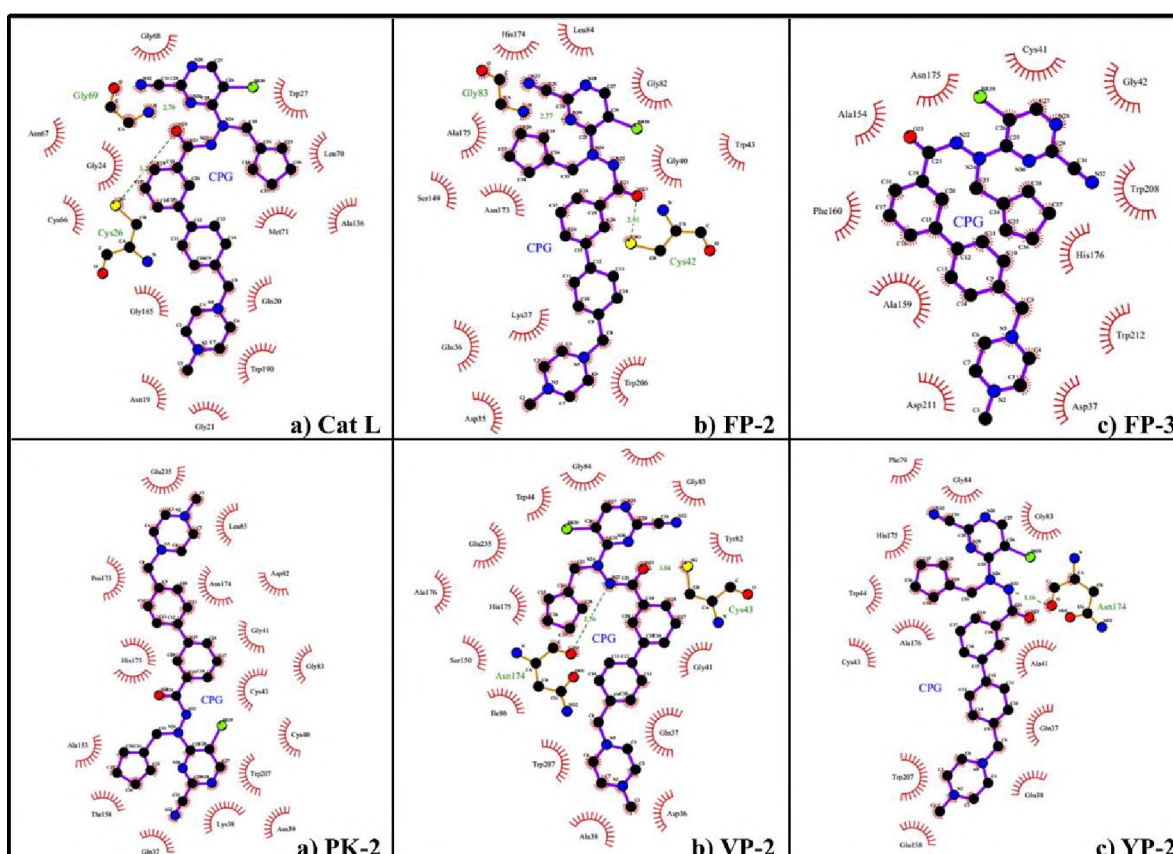


Figure 3.25: The residues interacting with CPG (blue), CPH (magenta) and CPI (green) in the binding pocket of human cathepsins and *plasmodial* proteases during the docking stage.

In addition, several hydrogen bonds were formed although in most cases they were not stable (Appendix 1K, Figure 3.26). To understand the differential binding profiles, it is important we understand their chemical structures in details. All CPs had a common central scaffold to which two sets of chemical groups (R1 and R2) were attached. Each set had three different

substitutions in which R1 was of varied lengths while R2 differed in terms of their cyclization (cyclopentyl and cyclohexyl). In all cases, compounds CPD, CPE and CPF established fewer vdW and hydrogen interactions compared to the rest of CPs, an observation that could be explained by the short length of R1 (see Figure 3.5). In contrast, CPG, CPH and CPI maintained vdW interactions with most subsite aa residues mainly because of their extended chemical nature of R1. It is well known that hydrogen bonds play critical roles in the stabilization of protein-ligand complexes. From H-bond analysis, the human cathepsins exhibited low unstable hydrogen bond occupancy compared to majority of the *plasmodial* homologs, a factor that can explain the differential docking energies observed (Appendix 1H).

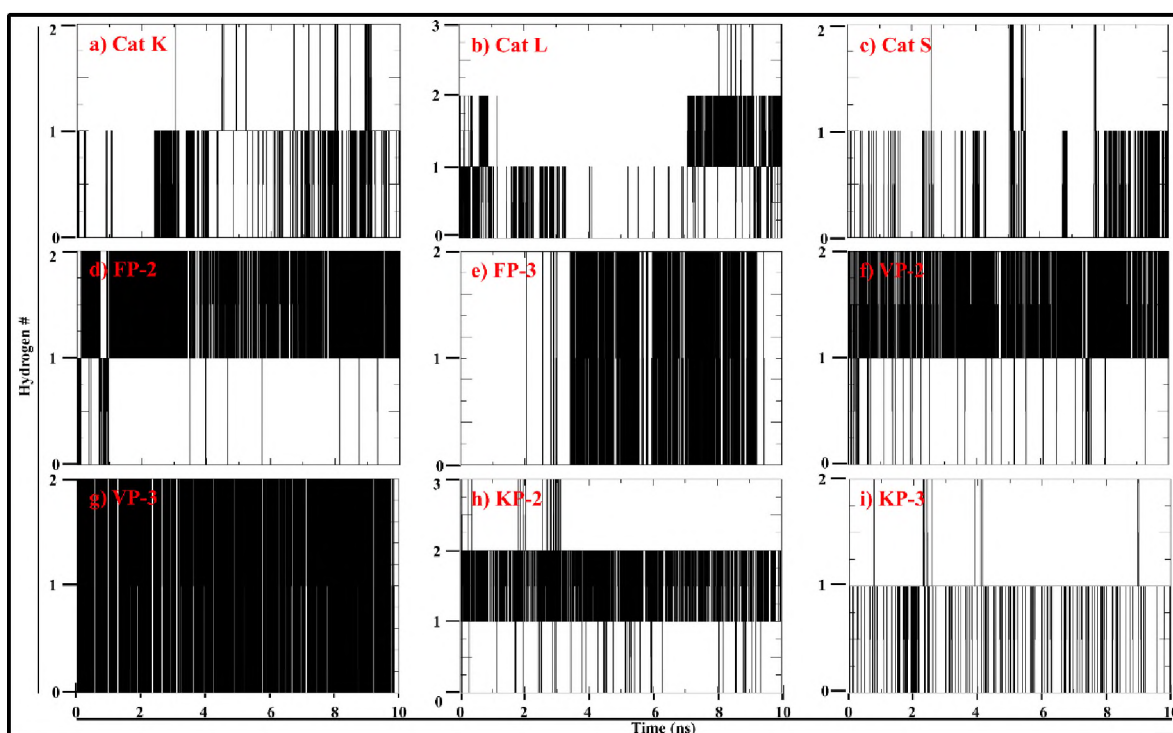


Figure 3.26: Hydrogen bond dynamic profiles of different protein homologs when in complex with CPG during the entire MD simulation period.

3.5.3.7 Conformational changes during simulation

Understanding the dynamic changes of both a ligand and protein is essential in drug design³²². To determine the conformational changes occurring throughout the simulations,

protein-ligand complexes were generated at 2 ns time interval and then visualised via PyMOL. Shown in Figure 3.27 is an example illustrating the conformational changes at specific time points of compound CPG in complex with Cat K (a) and FP-2 (b). To maintain interactions between a ligand and interacting residues, its conformation must be kept steady. From the results, CPG attains a stable conformation all through from 6 ns. Both the R1 and R2 chemical groups maintained stable interactions with the S1' and S2 subsites respectively. These results were observed in the other proteins when in complex with CPG-CPI. Also shown are the zoomed view of interacting residues and their corresponding subsites. A key observation was that CPG interacted with both proteases mainly via vdW and hydrogen interactions. Majority of the vdW interactions were as a result of S1' residues.

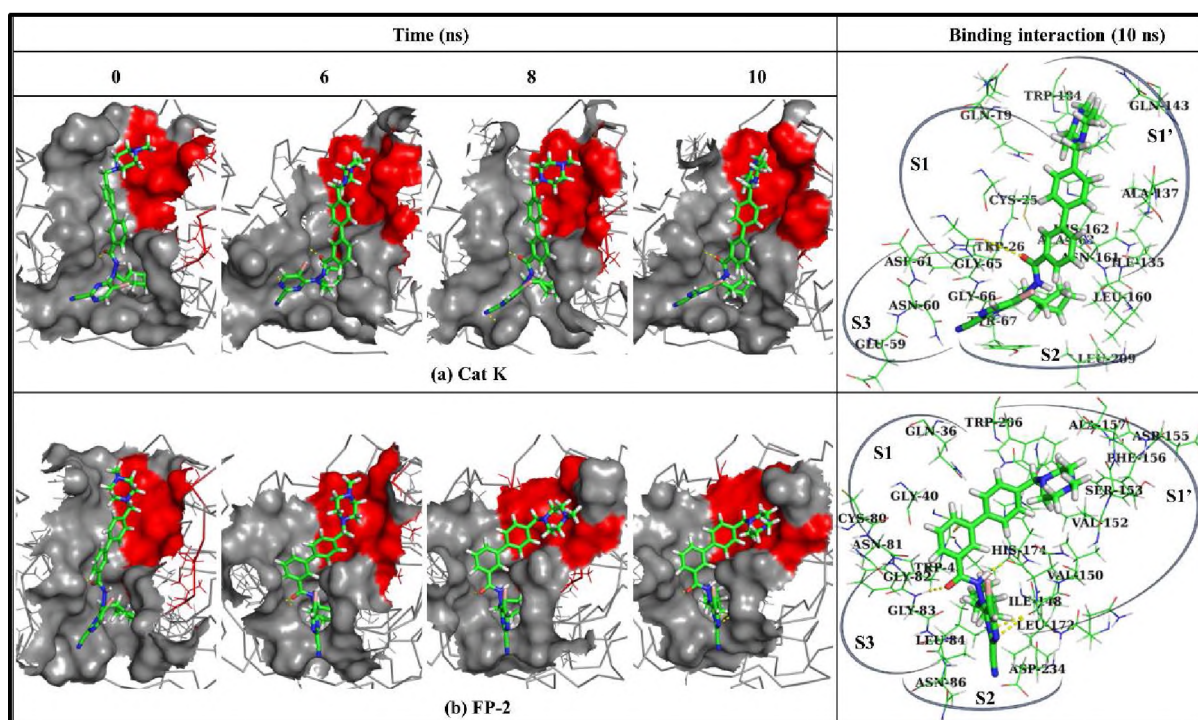


Figure 3.27: Ligand conformational changes over time. Dynamical evolution of compound CPG (ball and stick representation) when in complex with Cat K (a) and FP-2 (b) at the start (0 ns), 6 8 and 10 ns. Highlighted in red is the S1' subsite. The corresponding panels on the right depict the interacting residues with CPG either through vdW or hydrogen bonds (yellow dashes). Used in Musyoka TM *et al.*, 2015²²⁴.

3.5.4 5PGA and selected ZINC hits

Previous docking studies using a set of 23 compounds identified a single potential hit named 5 α -Pregna-1,20-dien-3-one (5PGA) from *Capnella thyrsoidea*. From literature, the sterol like compound had never been tested previously for antimalarial activity. However, a different study established that it had the ability to elicit inflammatory response in the neutrophils of rabbit cells via the release of superoxide ions³²³. To further study its potentiality as a potential hit for antimalarial drug discovery, it was used to search for similar compounds from the ZINC chemical database through the LBVS approach. This led to 186 similar compounds which were screened for activity against the *plasmodial* and human proteases. A filtering approach based on the docking energy results was applied and the compounds with lowest energy and with strong activity and selectivity profiles against multiple *plasmodial* and human cathepsins respectively selected. Ultimately, five analogs of 5PGA from ZINC were selected for further studies.

A total of 66 MD simulation runs each of 20 ns involving 11 proteins and six ligands *viz.* 5PGA (SA natural compound), ZINC36371307, ZINC03869631, ZINC04532950, ZINC04579000 and ZINC05247724 (ZINC hits) were performed. To determine the stability and mechanistic aspects of the protein-ligand interactions, the RMSD, RMSF and radius of gyration (Rg) of each system were determined. However, in comparison with the previous class of compounds, the determination of these observables was from the last 12 ns of every trajectory. The results of this section have been published (Musyoka *et al.*, 2016).

3.5.4.2 RMSD

To determine the global stability of the different protein-ligand systems throughout the entire simulation periods, the RMSD of their C α atoms for the apo, holo and ligand structures were determined with the reference being the initial structure. During the first 4 ns, the RMSD

values in all the systems dramatically increased to $\sim 0.22 \pm 0.2$ nm after which they started to converge (Figure 3.28).

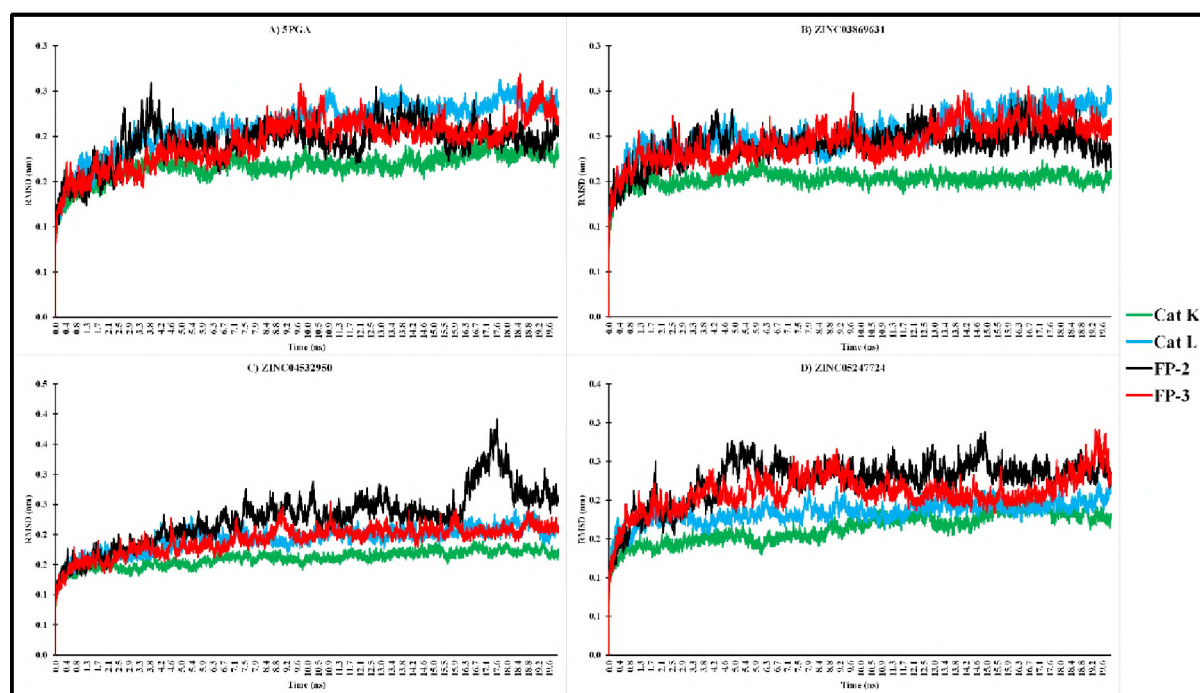


Figure 3.28: The RMSD plots of human cathepsins (Cat K and Cat L) and the FPs (FP-2 and FP-3) when in complex with a natural SA compound (5PGA) and selected analogs from the ZINC database during a 20 ns MD simulations.

Just like with CPs, the Apo forms of the human cathepsins had the least RMSD values with Cat K having a values of 0.10 ± 0.1 nm while Cat L 0.14 ± 0.1 nm. For the *plasmodial* proteases, the RMSD values were 0.18 ± 0.3 nm (Figure 3.29A). From Figure 3.29B, the process of ligand binding slightly increased the overall complex RMSD with Cat K having values of 0.16 ± 0.2 nm, Cat L of 0.20 ± 0.2 nm while the *plasmodial* proteases had values ranging between 0.23 ± 0.3 nm. All these results were in the same range as with the previous set of compounds (CPs).

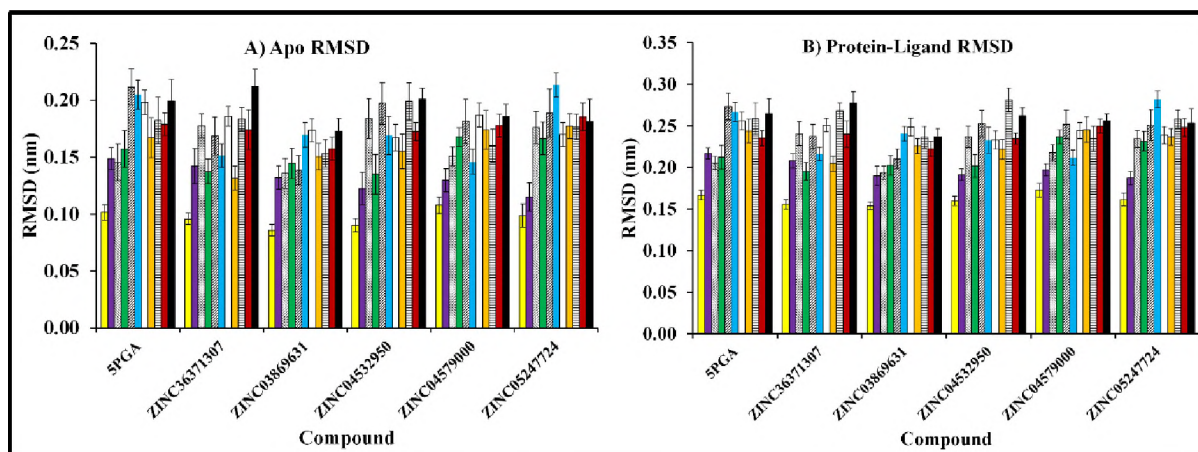


Figure 3.29: The average RMSD values of the apo (a) and holo (b) systems for the time period between 8 and 20 ns. Error bars indicate RMSD standard deviation per system over a 12 ns timescale.

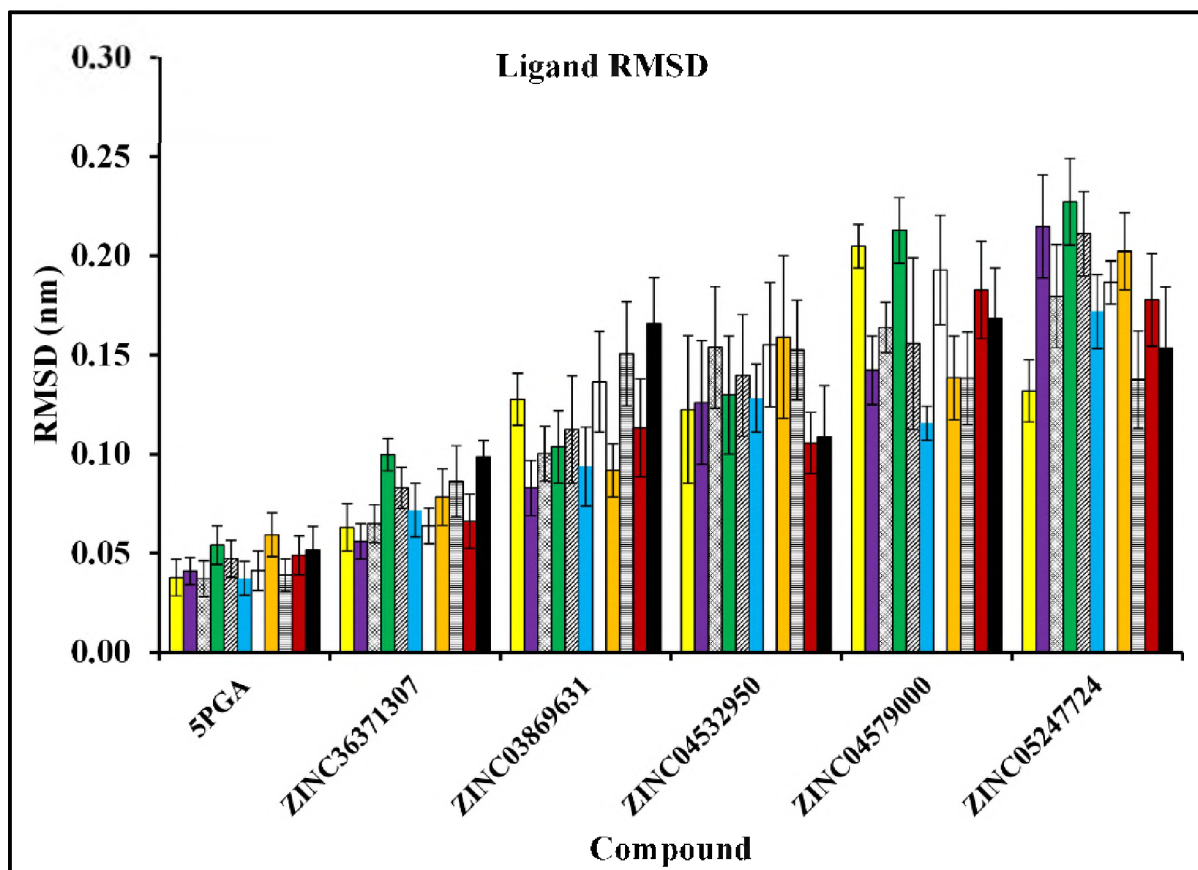


Figure 3.30: The average RMSD values of 5PGA and its ZINC analogs for the time period between 8 and 20 ns. Error bars indicate RMSD standard deviation per ligand over a 12 ns timescale.

From the ligand RMSD (Figure 3.30), both 5PGA and ZINC36371307 had the least RMSD values of ~ 0.05 and ~ 0.85 nm respectively. This was in relation to their planar chemical structure of and the lack of rotational bonds. In ZINC04532950, ZINC04579000, and

ZINC05247724, more fluctuations were observed a fact linked to the increased presence of rotational bonds within their structures. A comparison of the recorded ligand RMSDs and those of the respective holo structures together with the apo proteins, it can be concluded that ligand binding did not affect the proteins' overall conformational diversity significantly.

3.5.4.3 Rg

All the protein structures remained compact during the entire simulation as depicted by the Rg plot (Figure 3.31). From the results, Cat K had the lowest Rg of ~ 1.65 when in complex with all the compounds while Cat L had values in the range of 1.69 ± 0.2 nm. Both the human and rodent *plasmodial* proteases had very similar RMSD values.

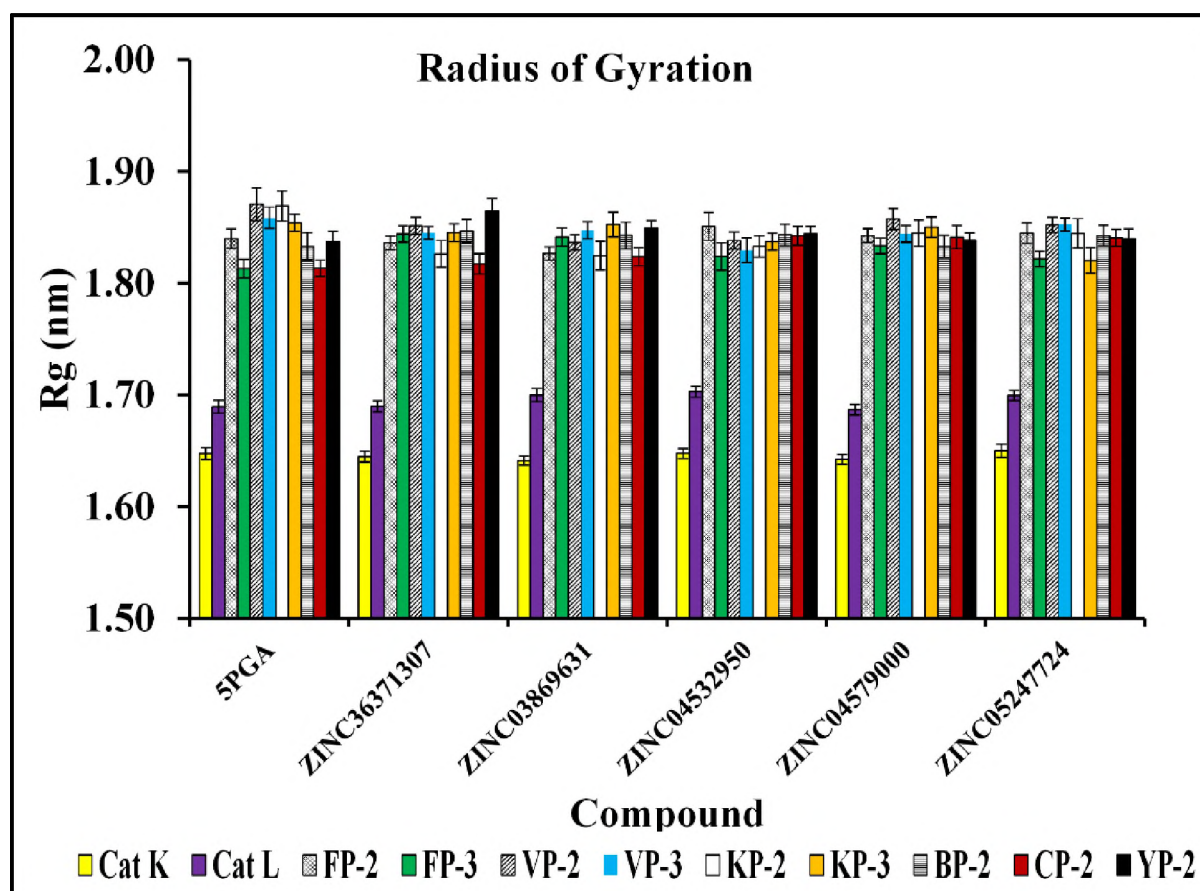


Figure 3.31: The average compactness of the different proteins when in complex with 5PGA and its analogs for the time period between 8 and 20 ns. Error bars indicate Rg standard deviation per ligand over a 12 ns timescale.

3.5.4.4 RMSF

To better understand the protein inherent flexibility, RMSF of the $\text{C}\alpha$ backbone was calculated. As was with the CPs set of compounds, greater fluctuations were recorded in the loop regions (Figure 3.32). *Plasmodial* proteases exhibited the largest flexibility around the characteristic inherent high fluctuating β -hairpin loop feature.

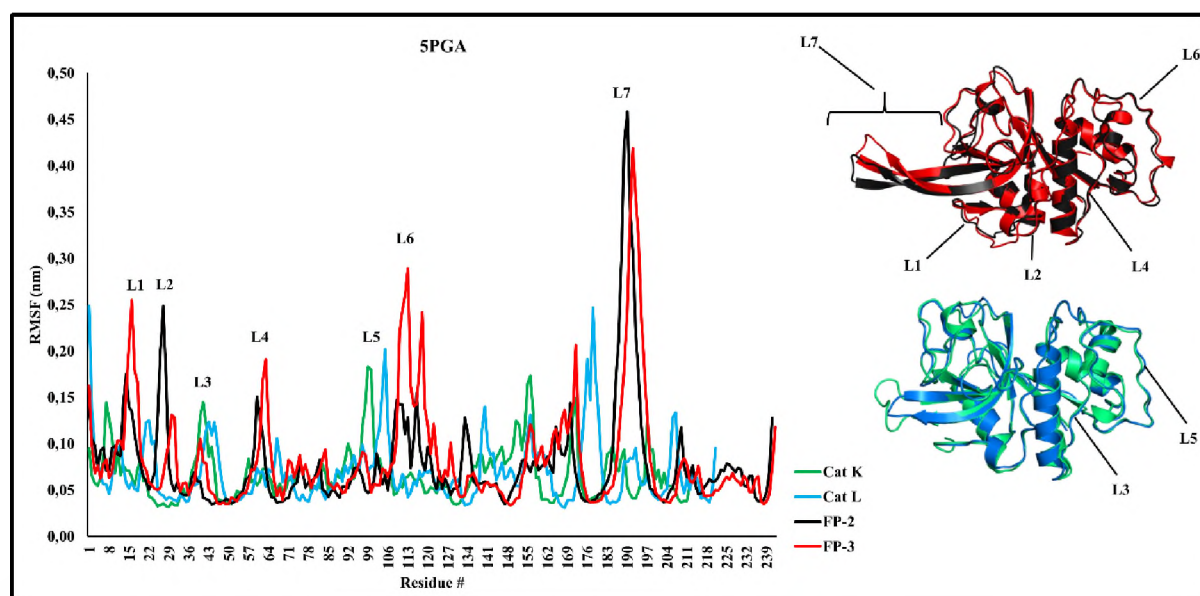


Figure 3.32: Local residue fluctuations of falcipains (red and black) and cathepsins (green and blue) when complexed with 5PGA during the last 12 ns of a MD simulations. Cartoon representation shows the different loop regions that are responsible for the observed high fluctuations. Color code corresponds to that of the RMSF plots.

3.5.4.5 Hydrogen bonding

The *g_hbond* tool in GROMACS was used to determine the number of hydrogen bonds and their occupancy. In most cases, the hydrogen bonds at the docking level (Table 3.3) were maintained during the MD simulations. An important feature to note was the differential bond occupancy between the human cathepsins and the *plasmodial* proteases (Figure 3.33 - 3.36). VMD was used to visualise hydrogen bond dynamics during simulation. Cathepsins exhibited lower hydrogen bond occupancy compared to the *plasmodial* homologs. For example, during the first 0.8 ns of MD simulation, a hydrogen bond formed between Gln143 and ZINC03869631 in Cat K was stable but thereafter the bond distance changed substantially leading to its breaking (Figure 3.33).

Table 3.4. A summary of interacting residues with the various ligands under study. Enclosed in brackets are residues forming H-bonds. Residue numbers are according to catalytic domains.

Cmpd/ Protein	5PGA	ZINC				
		36371307	03869631	04532950	04579000	05247724
Cat K	W184, (N187)	W26,Y67,A134, L160,H162, A163,L209	W26,W67,A134, A137,N161,H162, A163,W184, (Q143)	W26,Y67,A134,Q1 43,N161,H162,A163 ,W184, (Q143)	Q21,W26,Y67, A134,N161,H16 2,A163	Q21,W26,Y6 7,A134,,A163
Cat L	W27,L70, A136,M16 2	Q20,Q22,L14 5,H164,W190 ,W194	W27,L70,A136,A 139,D163,H164, W190	W27,L70,A136,M1 62,D163,H164,W19 0	Q22,L145,F146, H164,W190,W1 94, (Q20,H164)	W27,L70,A13 9,D163,H164, W190
FP-2	W43,L84,I 85,N173,A 175,D234, F236	W43,L84,I85, L172,H174,A 175,D234	W43,L84,I85,V15 2,L172,N173,H17 4,A175, (I85)	W43,L84,I85,V152, N173,H174,A175, W206, (I85)	N81,L84,I85,Q1 71,L172,N173, A175, (I85)	W43,L84,I85, V152,N173,H 174,A175,W2 06 (I85)
FP-3	W45,Y86, N175,A17 7 (I87)	W45,Y83,Y86 ,I87,P174,A17 7	W45,Y86,I87,P17 4,N175,H176,A17 7,W208, (I87)	W45,Y86,I87,P174, N175,H176,A177, W208, (I87)	W45,Y86,I87,P 174,N175,A177 ,W208, (I87)	W45,Y86,I87, P174,H176,A 177, (I87)
VP-2	W44,Y82, F85,I86,N 174,A176, E235 (I86)	W44,Y82,F85 ,I86,P173,N17 4,A176	W44,F85,I86,V15 3,P173,N174,H17 5,A176,W207, (I86)	W44,F85,I86,V153, P173,N174,H175,A 176, (I86)	W82,F85,I86,P1 73,N174,A176, W207,E235, (I86)	W44,F85,I86, V153,P173,N 174,H175,A1 76,W207,E23 5, (I86)
VP-3	Q36,N38, V157,W20 6,W210	N38,A152,V1 57,H174,W20 6,W210	W43,I85,A152,V 157,N173,H174,A 175,W206	W43,N84,I85,A152, V157,N173,H174,A 175,W206	W43,N84,I85,P 172,N173,H174 ,A175,Q234	Q36,N38,V15 7,H174,W206 ,K209,W210, (H174)
KP-2	W44,L85,I 86,N174,A 176 (I86)	W44,L85,I86, P173,N174,A 176	W44,L85,I86,P17 3,N174,H175,A17 6,W207,E235, (I86)	W44,L85,I86,P173, N174,H175,A176, W207,E235, (I86)	L85,I86,P173,N 174,A176,W207 ,E235, (I86)	W44,L85,I86, P173,N174,H 175,A176,207 ,E235, (I86)
KP-3	W42,F84, N148,T17 1,N172,A1 74 (I84, N148)	W42,F83,I84, N148,T171,N 172,A174	W42,F83,I84,N14 8,T171,N172,H17 3,A174,W205, (I84)	W42,F83,I84,N148, V151,T171,N172,H 173,A174,W205, (I84,N148)	W42,D80,F83,I 84,N148,T171, N172,A174, (I84)	W42,F83,I84, N148,T171,N 172,H173,A1 74, (I84)
BP-2	Q37,A41, E158,W20 7,W211	K39,A41,V15 3,E158,H175, W207,W211	Q37,A41,W44,L8 6,A150,V153,N17 4,H175,A176,W2 07	Q37,A41,W44,V15 3,N174,H175,A176, W207	Q37,A41,W44, V153,E158,N17 4,H175,A176,W 207	Q37,A41,W4 4,V153,N174, H175,A176, W207
CP-2	A41,W44, I85,L86,P 87,150A,N 174,A176, Q234,Y23 6	Q37,R39,A41, Q158,H175,W 207,W211	A41,W44,I85,L86 ,A150,F172,A173 ,N174,H175,A176 ,W207,Q234	A41,W44,I85,L86, A150,F172,A173,N 174,H175,A176,W2 07,Q234	Q37,A41,W44, L86,A150,A173 ,N174,A176,W2 06	Q37,A41,W4 4,A153,Q158, H175,W207
YP-2	Q37,K39, A41,V153, W207, (K39)	A41,W44,I85, L86,A150,Y1 72,A173,N17 4,A176,Q234	Q37,A41,W44,I8 5,L86V153,Y172, N174,H175,A176, W207	A41,W44,I85,V153, Y172,A173,H175,A 176,W207,Q234, (Q234)	Q37,A41,W44, V153,N174,H17 5,A176,W207	Q37,A41,W4 4,V153,N174, H175,A176, W207

At docking stage, no hydrogen bond was observed in 5PGA-Cat L complex. However, after 8 ns of MD simulations, 5PGA carbonyl oxygen changed orientation forming a weak hydrogen bond with Lys118 (non-subsite residue). In Cat K, 5PGA formed a hydrogen bond with

Asn187 (non-subsite aa) which had a high occupancy during MD simulations (Figure 3.33a). For ZINC03869631, a H-bond with Gln143 (S1' residue) of Cat K was on for less than 0.5 ns during equilibration while in Cat L, there was no H-bond formed during the entire simulation (Figure 3.34).

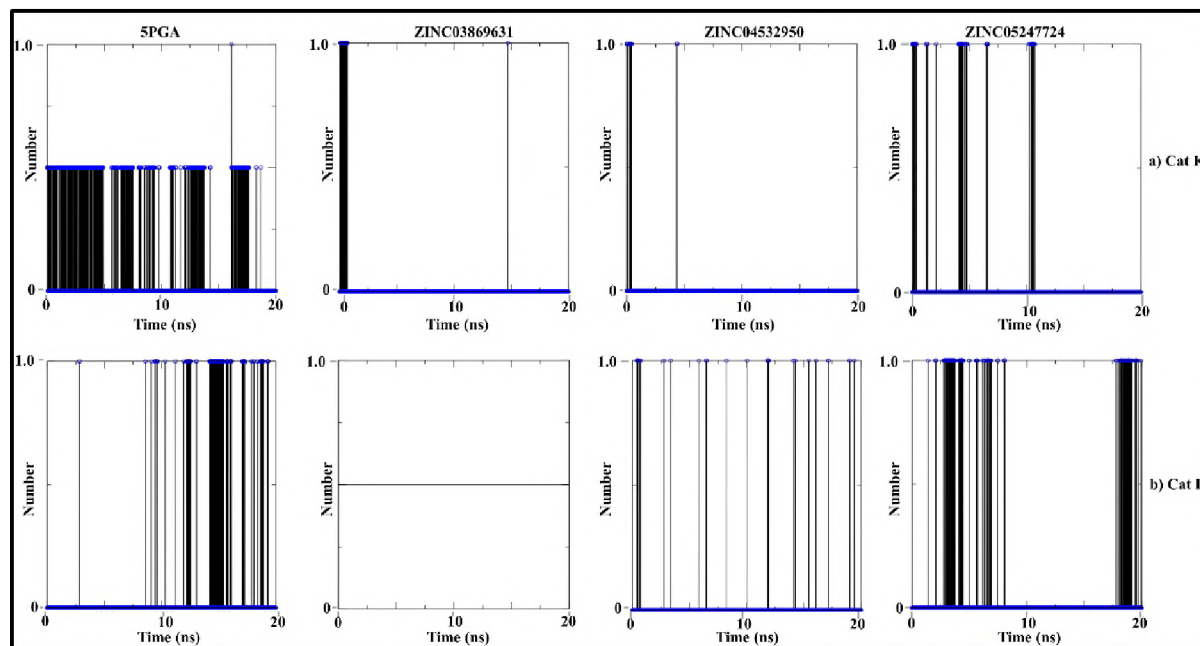


Figure 3.33: The number and evolution of intermolecular H-bonds of a) Cat K and b) Cat L in complex with 5PGA, ZINC03869631, ZINC04532950 and ZINC05247724 during a 20 ns MD simulation.

For ZINC04532950 and ZINC05247724, a similar trend of unstable H-bond formation was observed. However, in *plasmodial* proteases, the hydrogen bond occupancy was higher except in FP-2 when in complex with 5PGA. The hydrogen bonds formed between FP-2 and FP-3 with ZINC03869631, ZINC04532950 and ZINC05247724 showed greater stability (Figure 3.35).

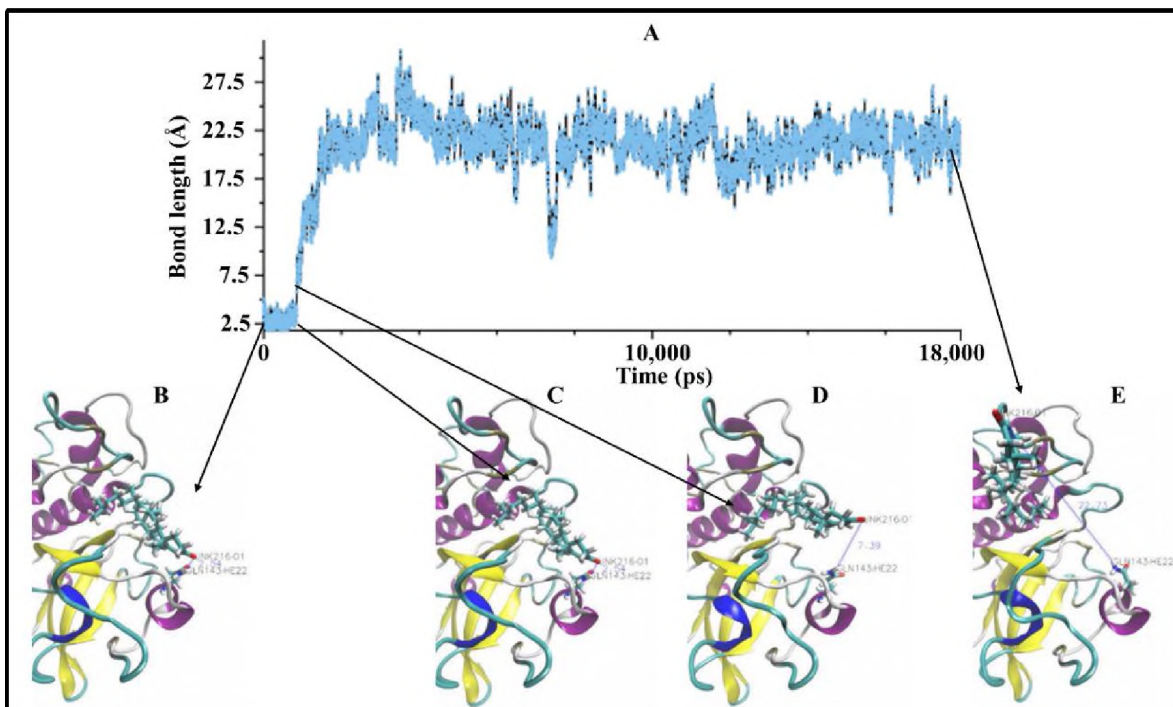


Figure 3.34: The evolution of H-bond length between Cat K Gln143 and ZINC03869631 and ligand orientation at different time points during MD simulation.

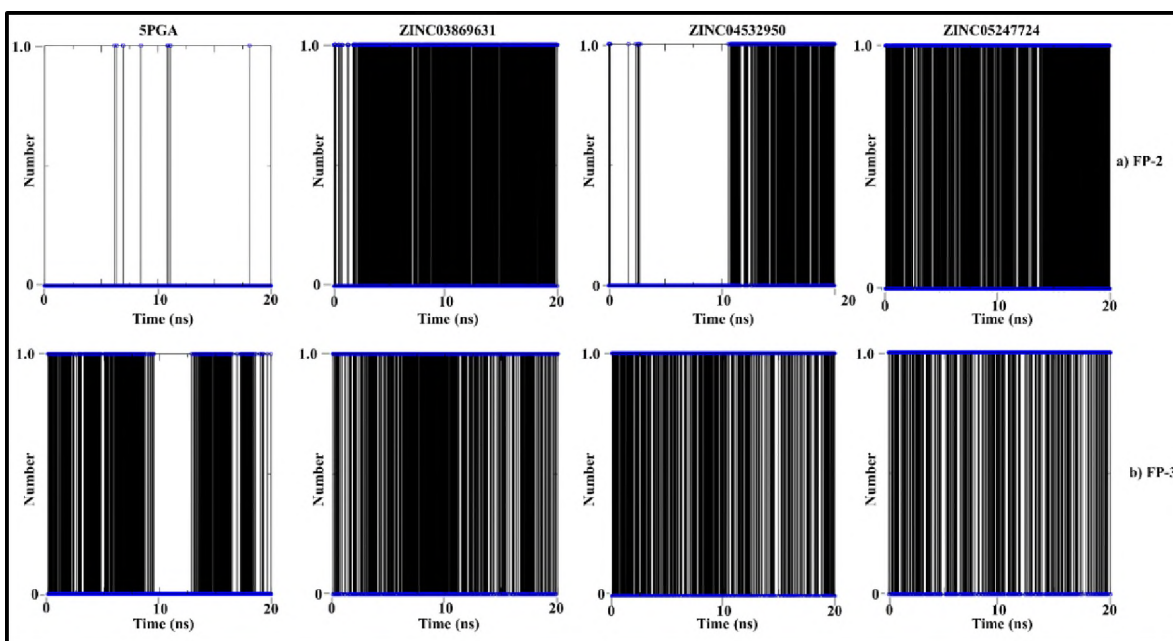


Figure 3.35: The quantitative and qualitative analysis of H-bonds of a) FP-2 and b) FP-3 in complex with 5PGA, ZINC03869631, ZINC04532950 and ZINC05247724.

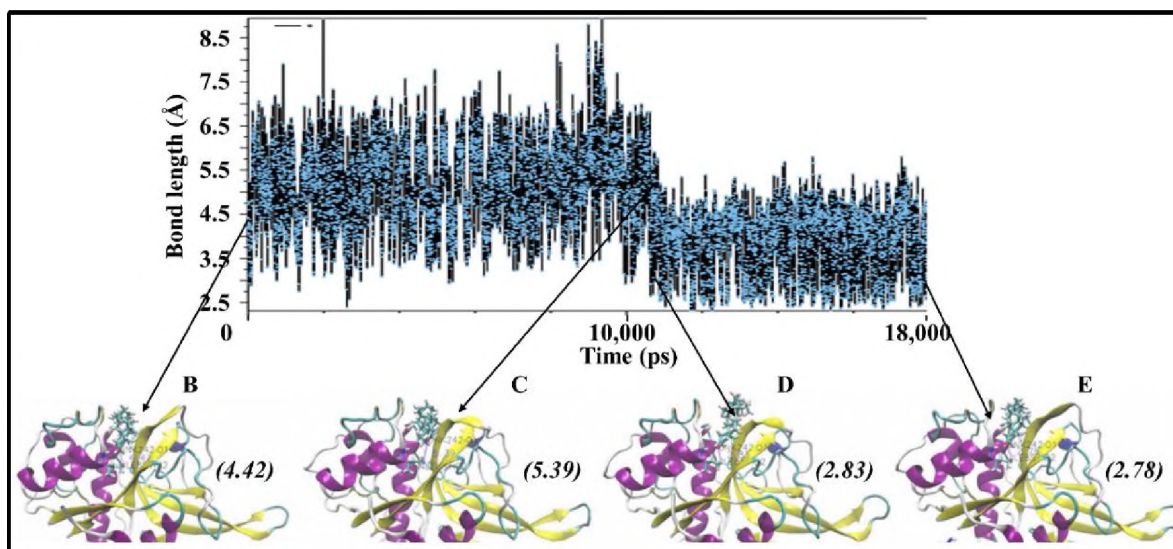


Figure 3.36: Evolution of H-bond length between FP-2 Ile85 and ZINC03869631 and ligand orientation at different time points during MD simulation.

Figure 3.36 shows the evolution of the only bond between FP-2 and ZINC03869631. The observed differences in hydrogen bond formation can be of valuable use in the design of further derivatives with better binding affinities and selectivity for the *plasmodial* and human cathepsins respectively.

3.5.4.6 Secondary structure element stability

DSSP algorithm was used to get some insights concerning the stability of the various protein-ligand systems by evaluating the number and changes in secondary structure during MD simulations. In all systems, there were no significant changes in structural elements observed during the entire simulation time (Figure 3.37). *do_dssp* tool in GROMACS 4.6.5 was utilised as GROMACS 4.5.5 is devoid of the program.

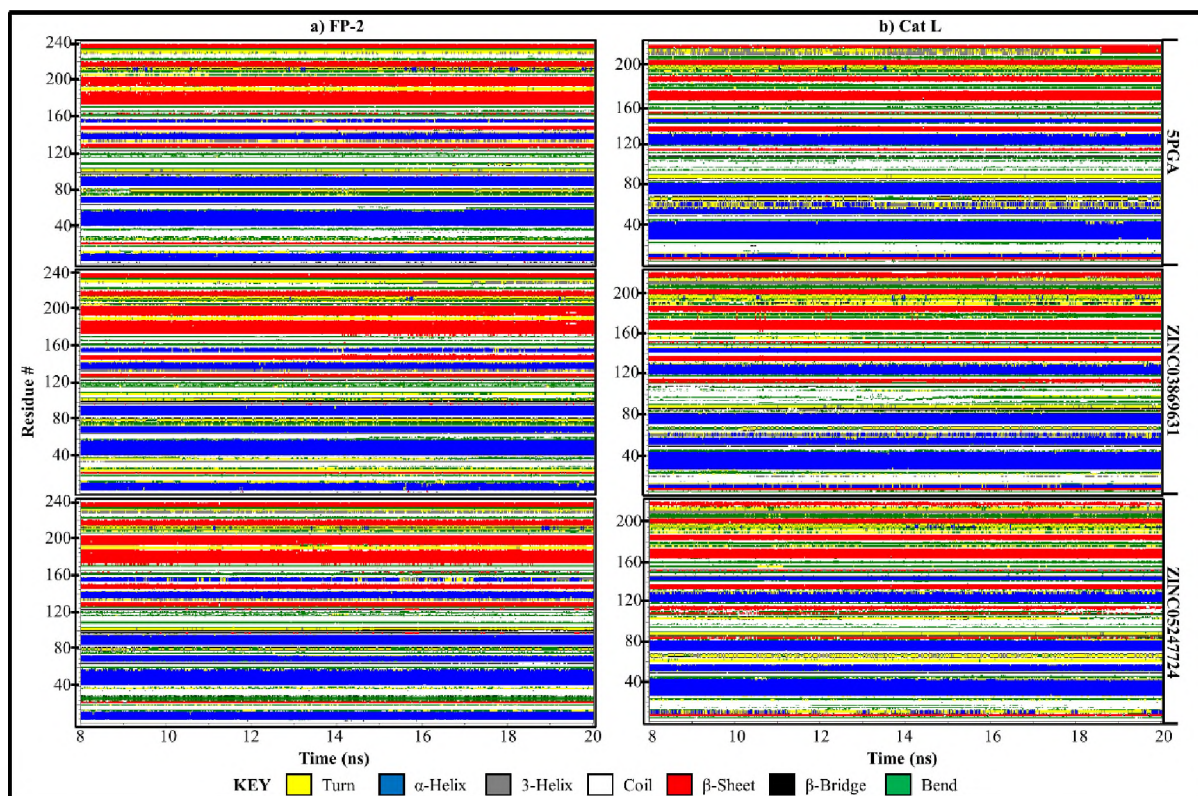


Figure 3.37: Conformational evolution of secondary structure elements of a) FP-2 and b) Cat L in complex with 5PGA, ZINC03869631 and ZINC05247724 during the last 12 ns of MD simulation as determined by *do_dssp* tool.

As seen with FP-2 and Cat L in association with 5PGA, ZINC03869631 and ZINC05247724, the helical and β -sheet content remained constant during the MD simulations (Figure 3.37). This further confirmed the stability of the proteins as previously determined by GROMACS observables. Similar results were obtained with other *plasmodial* proteases.

3.5.4.7 Binding mode

From the ligand binding results, aa residues critical process can be determined. Differential binding profiles of 5PGA and best ZINC hits (ZINC03869631, ZINC04532950, and ZINC05247724) between human Cat L and FPs (Figure 3.38) were observed. In Cat L, the ligands showed diverse binding poses in comparison with *plasmodial* proteases. In FPs, all the ligands consistently bound with the same pose with an exception of FP-2-5PGA. In Cat L, Trp27, Leu70 (S2), Ala136 (S2), Asp163 (S2) and His164 (S2) were the main residues

involved in ligand binding. In FP-2 and FP-3, Ile85 and Ile 87 (S2 residues) respectively participated in hydrogen bond formation with a oxygen atom present across the ligand cohort.

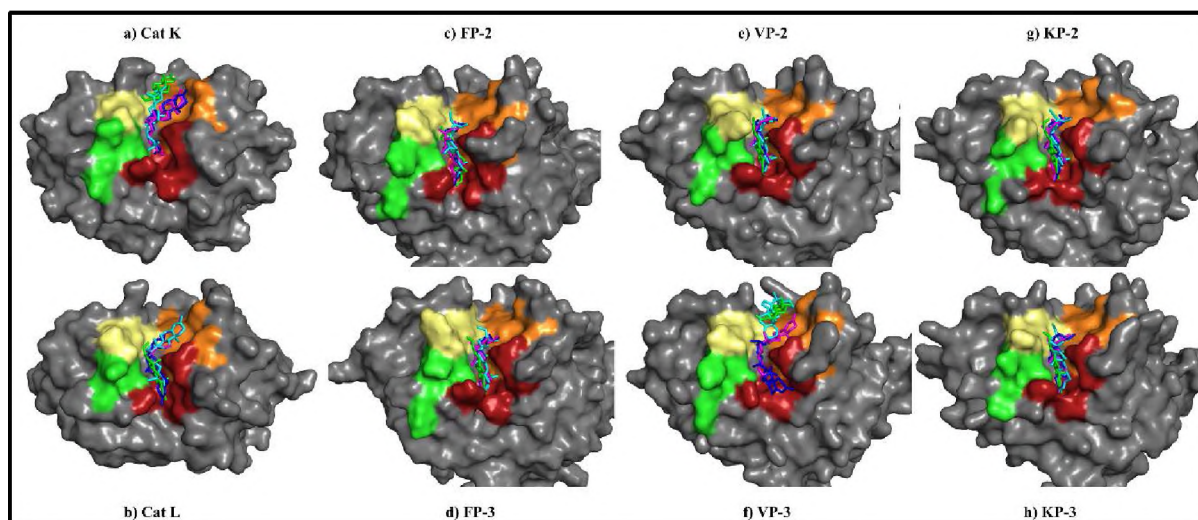


Figure 3.38: Binding poses of 5PGA (green), ZINC03869631 (magenta), ZINC04532950 (blue) and ZINC05247724 (cyan) in relation to the various subsites of cysteine proteases. S1 is shown in pale yellow, S2 in brick red, S3 in green while S1' in orange.

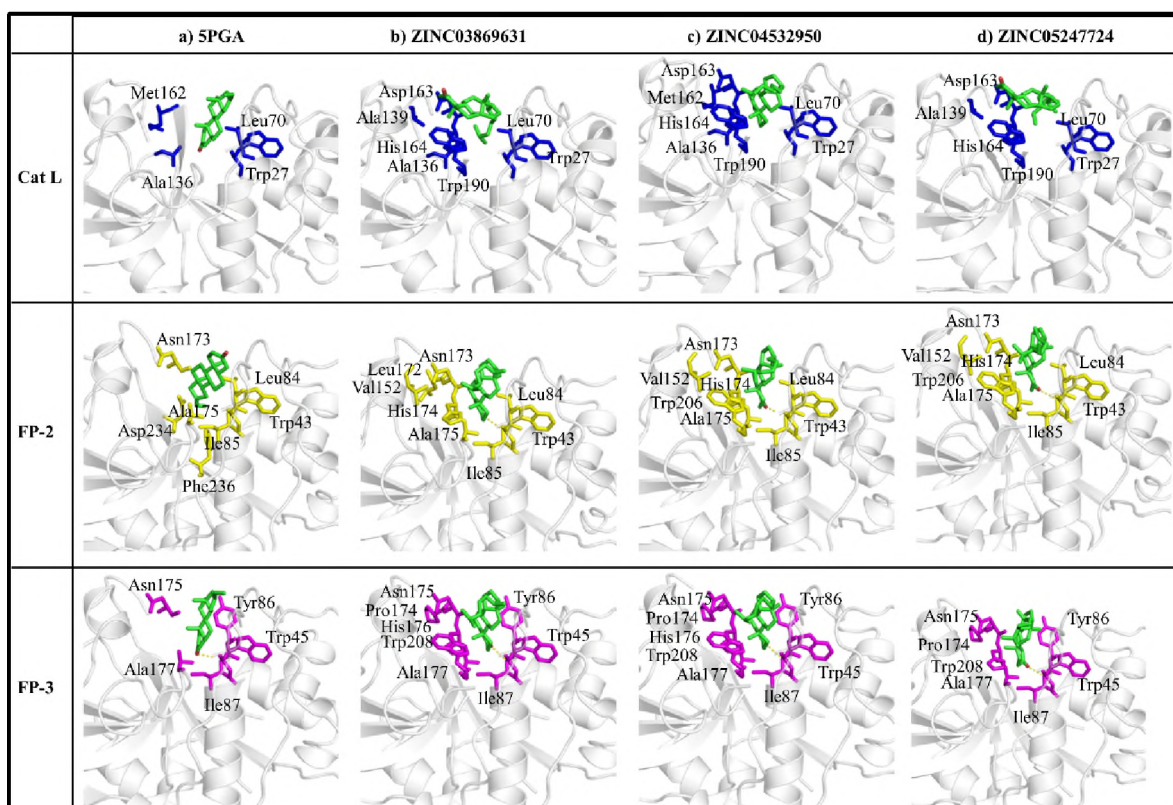


Figure 3.39: Binding pocket aa residue interactions patterns of bound 5PGA, ZINC03869631, ZINC04532950, and ZINC05247724 with Cat L (blue), FP-2 (yellow) and FP-3 (magenta). Hydrogen bonds are depicted in a yellow dotted line.

The main residues involved in binding of the different compounds were determined (Figure 3.39). These were the same residues as was with CPs although in the current set of compounds, fewer hydrophobic and hydrogens were observed. This was as a result of the smaller size and chemical nature of the current compounds.

3.5.5 Chemical modifications necessary for hit to lead compounds

Determination of drug-likeness of compounds has become an integral part of the drug development process especially during the early stages³²⁴. Majorly, this is to reduce the attrition rate at later stages as well as cut down the cost involved in the development process as only compounds with preferred pharmacokinetics properties are allowed to proceed with the process. By considering the physicochemical properties of a compound using *in silico* approaches, its molecular impact *in vivo* can be accessed which is mainly determined by its bioavailability and toxicity. Drug-likeness evaluation established that the identified hits violated some of the acceptable rules necessary for drug development (Table 3.4). CPs were the most drug-like compared to the rest of the hits studied. In the case of CPs CPG, CPH and CPI had slightly higher molecular weights than the acceptable (500 daltons). This resulted to higher octanol-water partition coefficient (LogP) and AlogP values which were within the acceptable range (<5). Despite having only one hydrogen bond donor (HbD), CPs had excellent number of hydrogen bond acceptors (HbA) thus explaining their high propensity in forming H-bonds with protein residues. For 5PGA and the ZINC hits, despite having acceptable molecular weights, their LogP values were higher than the acceptable value of 5. In addition, these compounds were devoid of hydrogen bond donors and had only one HbA. This explains their limited ability to form H-bonds with the protein residues. The net effect of this observation will be discussed in the next chapter. Compared to CPs, these compounds had comparable number of atoms. Thus, by comparing the molecular weight of these two sets of compounds, it can be noted that 5PGA and its analogs had higher number of H atoms.

Chemical modifications are necessary to replace these inert H-atoms with more functional groups. Hence, these results indicate that necessary chemical modifications must be effected in order to render the compounds important leads for further development of antimalarial drugs.

Table 3.5: Drug like properties of CPs, 5PGA and ZINC hits

Compound	Mwgt	LogP	AlogP	Property					
				HBA	HBD	TPSA	NoR	natom	TPSA
CPA	497.2	1.9	-1.3	8.0	1.0	87.3	7.0	60	87.3
CPB	483.1	1.4	-0.6	8.0	1.0	87.3	6.0	57	87.3
CPC	497.2	1.8	-2.4	8.0	1.0	87.3	6.0	60	87.3
CPD	489.1	1.7	-0.7	7.0	1.0	95.6	7.0	58	95.6
CPE	475.1	1.0	-1.6	7.0	1.0	95.6	6.0	55	95.6
CPF	489.1	1.8	-0.3	7.0	1.0	95.6	6.0	58	95.6
CPG	587.2	2.9	1.5	8.0	1.0	87.3	9.0	73	87.3
CPH	573.2	2.3	0.7	8.0	1.0	87.3	8.0	70	87.3
CPI	587.2	2.9	0.4	8.0	1.0	87.3	8.0	73	87.3
5PGA	298.2	7.4	1.6	1.0	0.0	17.1	1.0	52	17.1
ZINC36371307	424.4	11.8	3.2	1.0	0.0	17.1	1.0	79	17.1
ZINC03869631	384.3	10.5	2.2	1.0	0.0	17.1	5.0	72	17.1
ZINC04532950	370.3	10.0	1.5	1.0	0.0	17.1	5.0	69	17.1
ZINC04579000	410.4	11.1	1.9	1.0	0.0	17.1	5.0	76	17.1
ZINC05247724	412.4	11.6	2.0	1.0	0.0	17.1	6.0	78	17.1

3.5.6 MD pipeline

An advantage of *in silico* approaches is the ability to build up automated pipelines or workflows consisting of a diverse set of algorithms and subroutines that accept and process an input into results. With the introduction of grid computing and availability of high performance computer clusters, MD simulations can now be performed on a large scale. However, MD simulations are known to be complex and multistage processes that require prior knowledge in computer programming. To carry out a simulation successfully, one has to go through a rigorous process of system setup and execution which involves intensive manual control. This is a great challenge to experimentalists without any prior programming knowledge. Thus, a MD pipeline wrapped into a simple interface which allows MD neophytes to perform simulations in a automated mode is necessary.

Up to a total of 18 tools consisting of a set of in-house scripts and different GROMACS tools were wrapped into a single MD pipeline (Figure 3.40). This has since been established in the job management system (JMS) which is a workflow management system and web-based cluster front end for higher performance computing recently developed by RUBi³⁰¹. This is part of an ongoing plan to establish a fully integrated pipeline that will allow biologists to perform and analyse MD simulations with ease depending on the availability of computational resources. The only necessary requirements will be the availability of either an apo or holo structure and user defined parameters (force field, box dimensions, length of MD run etc) as these vary depending on the problem at hand. The MD pipeline is divided into seven major stages as shown in Figure 3.40 below. The user needs to have a 3D input of either a protein only (X-ray or homology model) or protein-ligand complex (docking experiments). By allowing the user to select different simulation parameters one can easily determine which of these parameters are more appropriate.

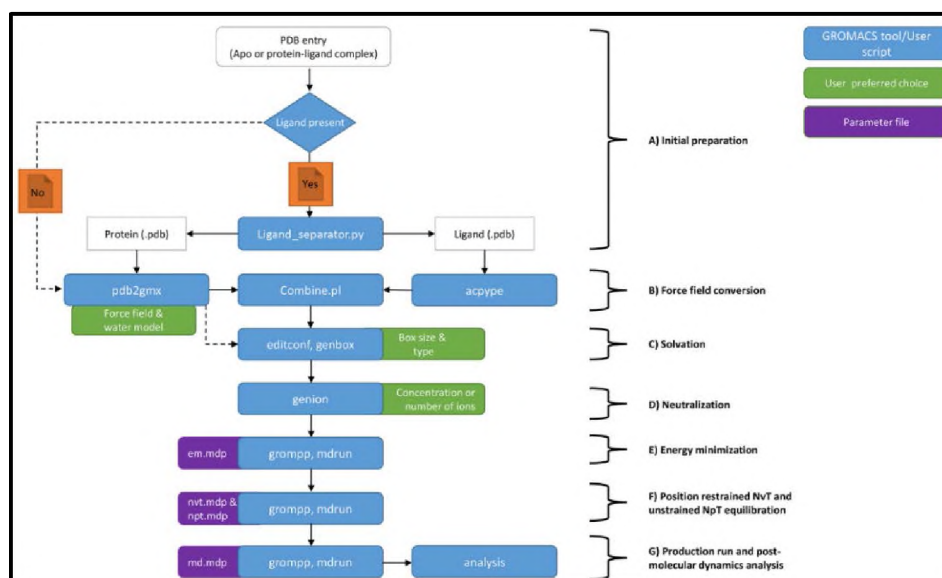


Figure 3.40: The automated MD simulation pipeline. The different GROMACS tools and ad hoc scripts (shown in sky blue) showing the flow of simulation steps during a classical MD experiment. Shown in orange are decision stages depending on the type of MD simulation (apo or holo) while in green indicates stages requiring user input and magenta being parameter files that determine how the MD runs will proceed. Used in Brown, DK *et al.*, 2015³⁰¹.

A user can quickly construct a sequential workflow of the different steps and indicate the dependency of each level as shown in Figure 3.41.

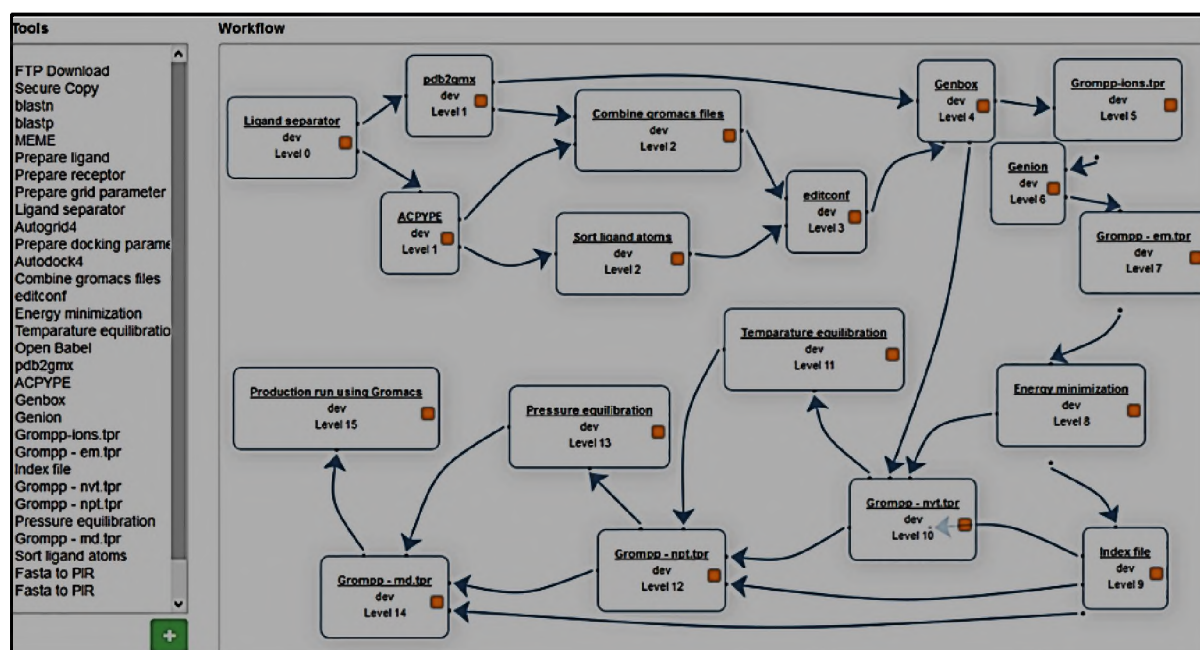


Figure 3.41: A diagrammatic representation of how the different GROMACS tools for the MD pipeline tool are linked and the dependencies of each step. The blue arrows indicate the sequential flow of the different MD process and the input/output requirement of each step (level). Used in Brown, DK et al., 2015³⁰¹.

Although this is still in the development phase, a major technical touch to improve its performance, efficiency and adaptability is needed. Currently, developments aiming to combine MD simulations, analysis and binding free energy (BFE) which will be dealt in the next chapter are ongoing. More details in regards to the future of the pipeline will be addressed in the concluding chapter (Chapter 6).

3.6 Chapter conclusion

The importance of MD simulations in the study of protein functioning cannot be underestimated. Proteins are dynamically fluctuating entities and thus experimental techniques like X-ray provide only a point of a high-dimensional configuration space that a protein could explore. Through MD simulation studies, conformational evolution of how the

different compounds studied was determined. Key factors that influence the binding of the ligands onto the different proteases were identified. These included ligand size, protein binding pocket size and chemical constituents of the ligand. This work emphasizes on the importance of MD studies in drug design. While docking data provides static interactions information between the compound and the receptor, it turns out to be inadequate in characterization of protein-ligand interactions. Although MD simulations offer a wealth of insight about the dynamical evolution of a protein-compound ligand complex, they lack accurate means to quantify the strength of association in the complex. Thus, the next chapter will focus on determining the strength of interactions between the complexes of each protein and ligand studied.

CHAPTER 4

Binding free energy calculations

The determination of binding free energy (BFE) between a protein and a ligand (substrate or inhibitor) is integral in understanding the principles that govern molecular recognition and conformational equilibrium. This is of central importance in medicinal chemistry and the pharmaceutical industry as free energy property determines the fate of a biophysical reaction. The past one decade has seen major advancements in the BFE determination methods. This is mainly due to the increase in computational power and the successful development of various methods for determining BFE with better efficiency and accuracy. Despite these developments, accurate determination of BFE still remains elusive. Here, BFE of association in all protein-compounds complexes (CPs, 5PGA and selected ZINC hits) studied in Chapter 3 will be determined. A comparison will be made with the docking energies and MD results to determine if there is any correlation in the three methods.

4.1 Protein molecular recognition

The existence of a cell is depended on array of chemical reaction processes with different paths³²⁵. Essential in regulation of these reactions are enzymes which present themselves as important drug targets³²⁶. During molecular recognition, proteins interact with a wide range of different other entities with a high degree of specificity and affinity³²⁷. To pharmacologically target these biomolecular structures, a complete understanding of their interactions is essential. Profiling the energetics between a protein and a ligand forms the basis of deciphering the molecular recognition processes. This involves qualitative and quantitative description of all the kinetics, thermodynamics and forces that govern formation of the specific molecular association. One characteristic feature with these biological catalysts is the occurrence of active sites which have steric and electrostatic features which enable ligand binding. It is important to characterise the energetic events that occur within this specific portion of the protein. As outlined by Copeland³²⁸, active sites have salient features such as their inherent volume always smaller than that of the whole protein, have a precise 3D arrangement of residues and cofactors with respect to that of the substrate, non-covalent interactions initiators of a binding event, located in a crevice or cleft for purposes of excluding solvent effects and the complementary structure of the ligand and the pocket shape determines their specificity. All these features have a contributory effect to the various processes and energetics involved in binding of a ligand.

An important application of understanding enzyme interactions is in the pharmaceutical industry. To achieve a therapeutic effect, a drug must stably interact with its target (usually a protein). Characterizing the strength of association as well as the structure of molecular complexes between a protein and a ligand offers invaluable insight in the structure-based of newer drugs of better efficacy and safety. During the drug development process, the determination of BFE plays a critical role in the identification of lead compounds. With the

continuous development of computational power and the improvement of various methodologies coupled with modelling tools, robust BFE studies can be achieved with reliable predictions. Several methods for determining the strength of protein-ligand interactions have been developed. These are broadly divided into experimental and computational approaches. This thesis is mainly focused on the latter.

4.2 Computational methods for BFE determination

Experimental techniques for BFE determination are quite expensive both financially and timewise. A better alternative is the use of computational approaches which have mainly been facilitated by the technical computational developments and better understanding of the binding theory³²⁹⁻³³¹. However, despite these developments, determining absolute BFE remains elusive mainly³³². This is majorly due to the failure to capture fully the translational, rotational and conformational entropies during molecular simulation which are also finite.

A range of computational approaches have been established to predict the binding affinities between a protein and a putative ligand. These range from more rigorous approaches that are based on molecular force fields to less rigorous methods that estimate BFE using simple energy functions³³¹. An important factor to determine when selecting the method for BFE analysis is the cost involved (computational and time). When performing virtual screening of large compound libraries, it is more practical to utilise less rigorous methods to identify hits that have desirable BFE profiles.

4.2.1 Empirical methods for BFE calculations

These methods use statistical scoring approaches based on simple energy functions or information from interactions between atom pairs³³³⁻³³⁵. Determine BFE by either counting the number of receptor atoms in contact with those of ligand or by calculating the change in solvent accessible surface area in the complex *vis a vis* of the uncomplexed protein and

ligand. As a result of using simplified energy functions as well as the treatment of solvent molecules in an explicit manner coupled with deficit of conformational sampling, these methods are fast but less accurate³³⁶. Include docking and several scoring functions have been developed such as Fscore, Xscore, Chemscore and FlexX function

4.2.2 Molecular force field methods for BFE calculations

These approaches which are considered rigorous and time consuming utilise MD or Monte Carlo simulation data to generate an ensemble of structures and determine the energy transformations occurring between different states³³⁶. Despite the lack of computational power, the statistical mechanical framework for BFE calculations together with several approximations were established a long time ago³³⁷. However, their application to biochemical systems was not until in the 1980's. The use of force fields in determination of BFE has been facilitated by the continued development of computational power owing to the expensive (computer resource) of this method. The first application of this method was reported in 1984 by Warshel, in which the energetics involved in proton transfer in lysozyme were determined³³⁸. At the same time, Tembe and McCammon described the binding of different ligands onto a receptor by the combination of free energy simulations coupled with thermodynamic cycles³³⁹. As there was an agreement between the computational and experimental results, these BFE methods became widely used. However, to achieve reliable and accurate results, there is a lot of efforts have been adopted to improve these methods³³⁶. Some rigorous methods include free energy perturbation (FEP) and thermodynamic integration (TI), umbrella sampling and potential of mean force (PMF).

4.3 Motivation

The work in this chapter is a continuation of the previous MD simulation studies (Chapter 3). As MD simulations lack accurate means to determine the energetics involved in binding, there is a need to determine both qualitatively and quantitatively the different forces involved

in binding of the identified hits. Here in, `g_mmpbsa`³⁴⁰, an open source BFE calculation tool that utilizes the molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) approach was used to determine the interaction energies between the various *plasmodial* and human proteases studied and the different hits (CPs, 5PGA and selected ZINC compounds). MM-PBSA, an attractive AMBER method originally developed by Kollman in the late 90s determines the strength of interaction between reagents and final product. It was first used to study the stability of RNA and DNA fragments³⁴¹. However, it has been modified in the recent years to determine the free energy of protein-ligand complexes based on the analysis of MD trajectories using a continuum solvent approach is gaining phase.

4.4 Methodology

Using the `g_mmpbsa` tool version 1.6, strength of interactions between different set of protein-ligand complexes was evaluated. A requirement of `g_mmpbsa` is availability of Adaptive Poisson-Boltzmann Solver (APBS) which determines the solvation energy properties of a system and GROMACS.

4.4.1 Preparation of input files

For successful BFE analysis, `g_mmpbsa` requires three input files namely; 1) a trajectory file which contains system snapshots during a MD simulation 2) an index file that contains the group of interest (in this case protein and ligand atoms in the system) and a parameter file containing necessary conditions and controls adopted for analysis. Pre-processing of the trajectory file to remove any periodic boundary conditions that results to unnecessary system jumps is necessary. Herein, this step was accomplished previously during the MD analysis step. An example of a parameter file is shown in Appendix 2G.

4.4.2 Executing `g_mmpbsa`

`g_mmpbsa` is a command based console application. For purposes of automation, a Bash script per every protein-ligand system was generated using a Python script, `pbsa_automate.py` (refer to Appendix 2H). A total of 174 runs; 108 for CPs (12 proteins and nine compounds), 11 for 5PGA (11 proteins and one compound) and 55 for ZINC hits (11 proteins and five compounds) were studied. Only the stable trajectory sections after equilibration were considered for the BFE calculations. A total of 4,000 and 6,000 system snapshots for each complex with the CPs and 5PGA together with its ZINC analogs were used for the BFE calculations respectively. Using the single trajectory approach the BFE (ΔG_{bind}) of each system snapshot was determined as follows:

$$\Delta G_{bind} = \Delta G_{complex} - (\Delta G_{receptor} + \Delta G_{ligand}) \quad (4.1)$$

where the free energies of the holo-protein, apo-protein and ligands are denoted by $\Delta G_{complex}$, $\Delta G_{receptor}$ and ΔG_{ligand} respectively. These are obtained by averaging of all snapshot geometries in a MD trajectory. In a gas-phase together with the solvation energy (G_{solv}) and entropy term (TS), the free energy (G) of each state was calculated as follows:

$$G = E_{gas} + G_{solv} - TS \quad (4.2)$$

T is system temperature which was set at 300 K. The gas phase energy (E_{gas}) is a sum of bonded (E_{int}) and nonbonded terms (E_{vdw} and E_{ele}) as shown in equation 4.3 below. Internal energy (E_{int}) comprises of contributions from bond, angle and torsion energies (equation 4.4).

$$E_{gas} = E_{int} + E_{vdw} + E_{ele} \quad (4.3)$$

$$E_{int} = E_{bond} + E_{angle} + E_{torsion} \quad (4.4)$$

$$G_{solv} = G_{pol} + G_{np} \quad (4.5)$$

$$G_{np} = \gamma SASA + b \quad (4.6)$$

The solvation term (G_{solv}) is composed of a polar solvation (G_{pol}) and nonpolar (G_{np}) energy components (equation 4.5). Polar solvation energies were determined by solving the Poisson-Boltzmann linear equation while nonpolar solvation through the solvent accessible surface area with an offset value (b) of 3.84928 kJ.mol⁻¹ and surface tension proportionality (γ) set at 0.0226778 kJ.mol⁻¹.Å⁻² (equation 4.6).

4.4.3 Analysis

To decipher the contribution of each energetic term involved in ligand binding process, the overall BFE term was decomposed to its individual components namely vdW forces, electrostatic energy, as well as polar and non-polar solvation energy. This was attained using a Bash script, BFE_decomposition.sh (Appendix 2I). A detailed per-residue decomposition

analysis to determine key protein residues contributing to the three components of BFE was also determined. Using a set of *ad hoc* batch scripts, the BFE runs together with corresponding analysis were automated.

4.4.4 System specifications

As the process of BFE analysis is highly expensive in terms of computational resources, all energy calculations were performed on the Tsessebe cluster (Sun) at the Centre of High Performance Computing (CHPC) Unit in Cape Town, SA.

4.5 Results and discussion

To determine how strong the interactions, the BFE of 108 CPs, 11 5PGA and 55 ZINC complexes was evaluated. In addition, the individual contributions by the various BFE energetic terms and aa were calculated to get a detailed picture of factors influencing ligand binding process. The energetic terms included vdW contributions, electrostatic (ele) interactions, polar solvation (PB) and entropy (SASA).

4.5.1 Role of BFE terms in the ligand binding process

In all the complexes studied, binding process was majorly favoured by vdW and electrostatic interactions while polar solvation impaired it. For vdW, this was acceptable as the binding pocket of the proteins being studied majorly consist of hydrophobic residues as seen in Chapter 2. In the case of electrostatic interactions, hydrogen bonds between the ligand and the receptor are the main contributors. Thus the propensity of different ligands to form hydrogen bonds as determined by number of hydrogen bond donors and acceptors mainly controlled this energetic term (Table 3.4). A summary of the overall BFE underlying the binding of CPs, 5PGA and its ZINC analogs to FP-2 and FP-3 as well as their homologs is shown in Table 4.1 and Table 4.2 respectively.

In the case of CPs (Table 4.1), the overall interaction energies followed a similar trend as the docking results²²⁴. Compound CPG, CPH and CPI had the lowest binding energies in most cases, an indication of stronger interactions compared to the other CPs. From the different energetic contributions (Figure 4.1 and Appendix 1L), it was evidently clear that vdW and electrostatic energetic components enhanced binding of the ligands while the polar solvation impaired it. From MD studies, these compounds exhibited highest number of vdW and hydrogen bonds when compared to the rest (Chapter 3). The nonpolar solvation energies which correspond to the burial of solvent-accessible-surface area (SASA) upon binding enhanced the binding process in all complexes were almost of the same order. As it was in the case with docking and MD studies, both *plasmodial* proteases and human cathepsins bound the CPs with comparable affinities. In the case of Cat L, it exhibited the strongest interactions with the compounds except with CPC. From the BFE decomposition results, vdW interactions were weaker in this case compared with the rest of the compounds thus the observed differential binding result.

Table 4.1: Protein-CP complexes overall binding free energy (ΔG_{bind}) in $\text{kJ}\cdot\text{mol}^{-1}$ as determined by `g_mmpbsa` tool.

Protein	Compound								
	CPA	CPB	CPC	CPD	CPE	CPF	CPG	CPH	CPI
FP-2	-84.9±0.2	-71.8±0.3	-80.8±0.2	-91.1±2.7	-93.1±0.2	-80.5±0.2	-131.5±0.2	-103.4±0.2	-99.6±0.2
FP-3	-77.0±0.2	-67.1±0.2	-66.7±0.2	-74.2±0.2	-74.7±0.2	-102.8±0.4	-87.6±0.2	-111.7±0.2	-105.7±0.3
VP-2	-95.0±0.2	-78.7±0.2	-60.6±0.2	-85.8±0.2	-72.4±0.6	-77.7±0.3	-116.2±0.3	-81.8±0.2	-93.0±0.2
VP-3	-112.4±0.2	-72.6±0.2	-59.0±0.5	-98.6±0.3	-55.0±0.2	-62.1±0.3	-104.8±0.2	-93.7±0.2	-85.7±0.2
PK-2	-115.5±0.3	-68.8±0.3	-68.8±0.4	-92.8±0.3	-99.0±0.2	-85.7±0.3	-129.2±0.2	-61.9±0.3	-86.8±0.2
PK-3	-80.3±0.4	-63.0±0.2	-81.7±0.3	-82.1±0.2	-80.4±0.2	-80.9±0.2	-68.5±0.3	-76.2±0.3	-80.9±0.3
BP-2	-92.9±0.2	-92.1±0.4	-82.8±0.4	-133.3±0.5	-80.5±0.5	-85.0±0.3	-135.4±0.4	-131.7±0.3	-94.9±0.3
CP-2	-73.0±0.2	-85.4±0.2	-71.9±0.1	-108.3±0.3	-122.6±0.3	-83.7±0.3	-103.9±0.4	-103.4±0.3	-98.1±0.3
YP-2	-91.6±0.3	-67.0±0.3	-82.1±0.2	-103.0±0.2	-114.6±0.3	-89.1±0.3	-92.3±0.2	-104.8±0.2	-97.8±0.3
Cat S	-97.7±0.2	-72.2±0.2	-83.9±0.3	-85.8±0.3	-85.6±0.2	-84.9±0.3	-93.8±0.3	-98.2±0.3	-81.8±0.3
Cat K	-86.4±0.3	-91.3±0.2	-93.6±0.2	-90.8±0.3	-104.1±0.2	-76.6±0.3	-96.3±0.3	-80.0±0.2	-81.8±0.2
Cat L	-132.4±0.3	-99.2±0.2	-44.7±0.5	-129.2±0.3	-117.5±0.2	-111.8±0.3	-117.3±0.3	-149.7±0.3	-147.6±0.2

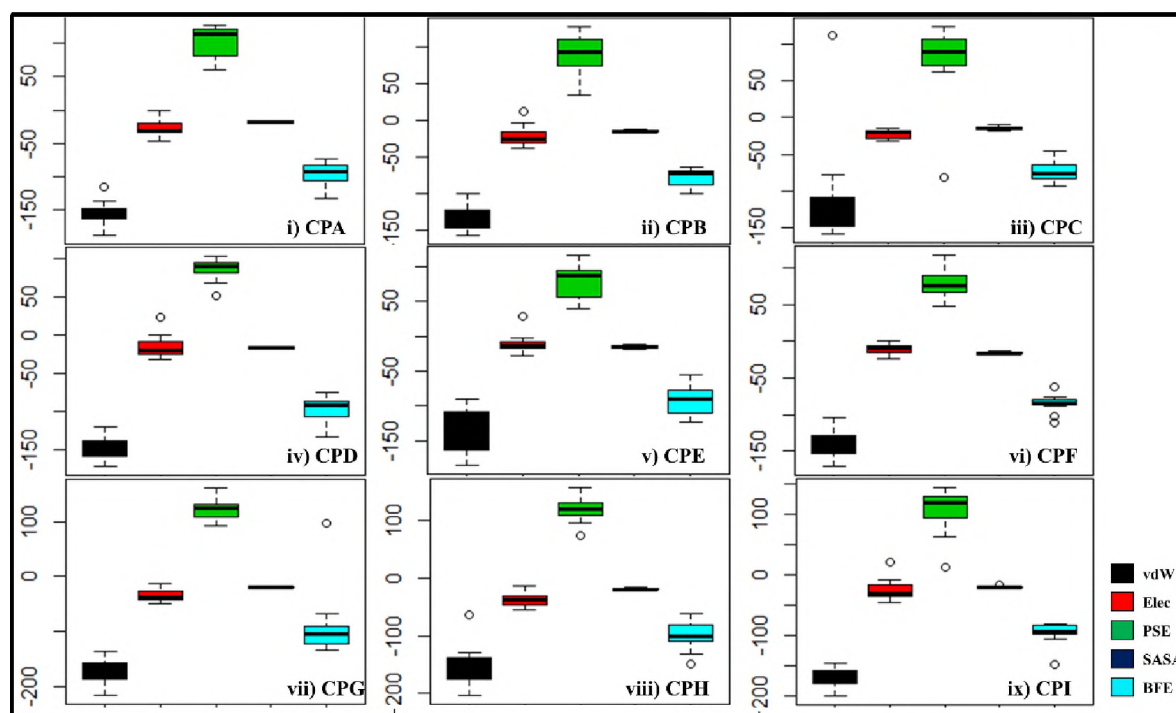


Figure 4.1: Box plots showing the distribution of the various BFE terms of CPs when bound to different proteins.

For 5PGA and its analogs, their BFE followed similar trend as with docking where ZINC hits exhibited stronger interactions compared to 5PGA (Table 4.2 and Appendix 1M). A detailed explanation of this observation will be discussed in the next section.

Table 4.2: The overall binding free energy (ΔG_{bind}) in $\text{kJ}\cdot\text{mol}^{-1}$ of the various proteases with 5PGA and selected ZINC compounds as determined by g_mmpbsa tool.

Protein	Compound					
	5PGA	ZINC36371307	ZINC03869631	ZINC04532950	ZINC04579000	ZINC05247724
Cat K	-78.6 ± 0.2	-113.0 ± 0.2	-99.2 ± 0.1	-91.8 ± 0.2	-81.3 ± 0.2	-93.4 ± 0.2
Cat L	-93.2 ± 0.2	-88.4 ± 0.1	-96.7 ± 0.2	-136.7 ± 0.2	-99.4 ± 0.2	-125.0 ± 0.2
FP-2	-86.7 ± 0.1	-81.2 ± 0.1	-86.7 ± 0.2	-83.8 ± 0.2	-70.6 ± 0.2	-80.1 ± 0.2
FP-3	-62.1 ± 0.2	-93.1 ± 0.2	-91.4 ± 0.2	-96.7 ± 0.2	-102.3 ± 0.2	-92.1 ± 0.2
VP-2	-65.3 ± 0.2	-101.7 ± 0.2	-91.3 ± 0.2	-92.3 ± 0.2	-81.6 ± 0.2	-107.9 ± 0.2
VP-3	-71.5 ± 0.1	-78.2 ± 0.2	-66.8 ± 0.2	-73.9 ± 0.2	-121.7 ± 0.7	-101.7 ± 0.2
KP-2	-68.2 ± 0.2	-109.9 ± 0.2	-91.4 ± 0.2	-81.9 ± 0.2	-68.0 ± 0.2	-71.5 ± 0.1
KP-3	-44.5 ± 0.2	-81.8 ± 0.2	-68.2 ± 0.2	-55.3 ± 0.2	-70.6 ± 0.1	-69.8 ± 0.2
BP-2	-65.6 ± 0.2	-90.1 ± 0.2	-92.7 ± 0.2	-84.0 ± 0.2	-74.2 ± 0.2	-82.7 ± 0.2
CP-2	-75.4 ± 0.1	-89.8 ± 0.2	-100.2 ± 0.2	-103.0 ± 0.2	-119.1 ± 0.3	-91.8 ± 0.2
YP-2	-62.9 ± 0.2	-76.6 ± 0.2	-84.7 ± 0.1	-96.7 ± 0.2	-86.4 ± 0.2	-79.3 ± 0.1

As was with the CPs, vdW and electrostatic contributions favoured the binding of 5PGA and its analogs (Figure 4.2). However, the resulting contributions were much lesser a fact that can be linked to their chemical nature and size (discussed in the next section).

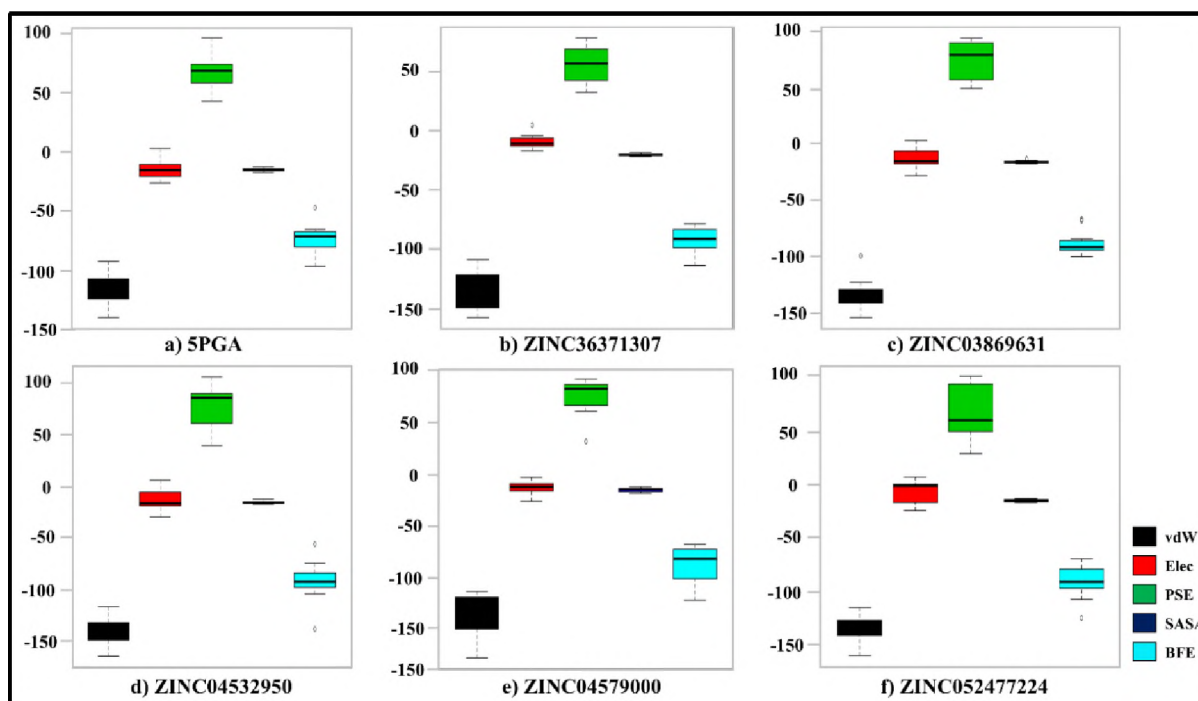


Figure 4.2: Box plots showing the distribution of the various energy terms of the different proteins when bound to 5PGA and its ZINC analogs.

4.5.2 Subsites contribution to BFE

To determine the important residues influencing the strength of interactions in each of the protein-ligand complexes, the final BFE per each discrete system was further decomposed into individual residue contributions. Figure 4.3 and 4.4 and Appendix 1N show key residues that influence the BFE either positively or negatively for various proteases when complexed with CPs (CPG, CPH and CPI), 5PGA and ZINC hits (ZINC03869631, ZINC04532950 and ZINC05247724).

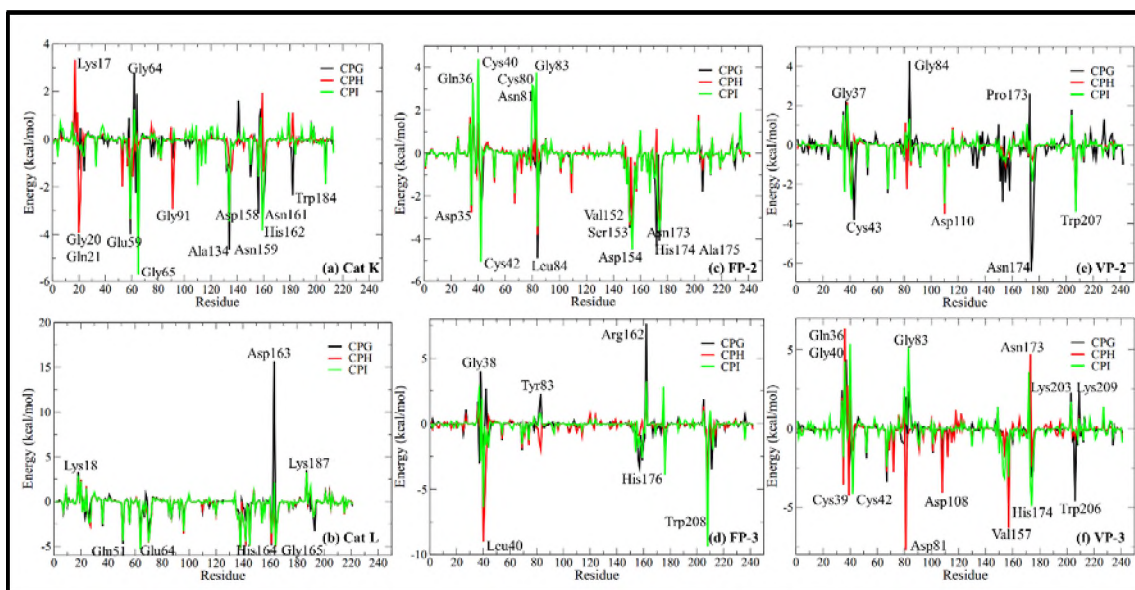


Figure 4.3: A detailed per-residue fingerprint showing the individual aa energy contributions to the binding of CPG (black), CPH (red) and CPI (green) with (a) Cat K (b) Cat L (c) FP-2 (d) FP-3 (e) VP-2 and (f) VP-3. Positive values indicate residues that impair binding and vice versa. Used in Musyoka TM *et al.*, 2015²²⁴.

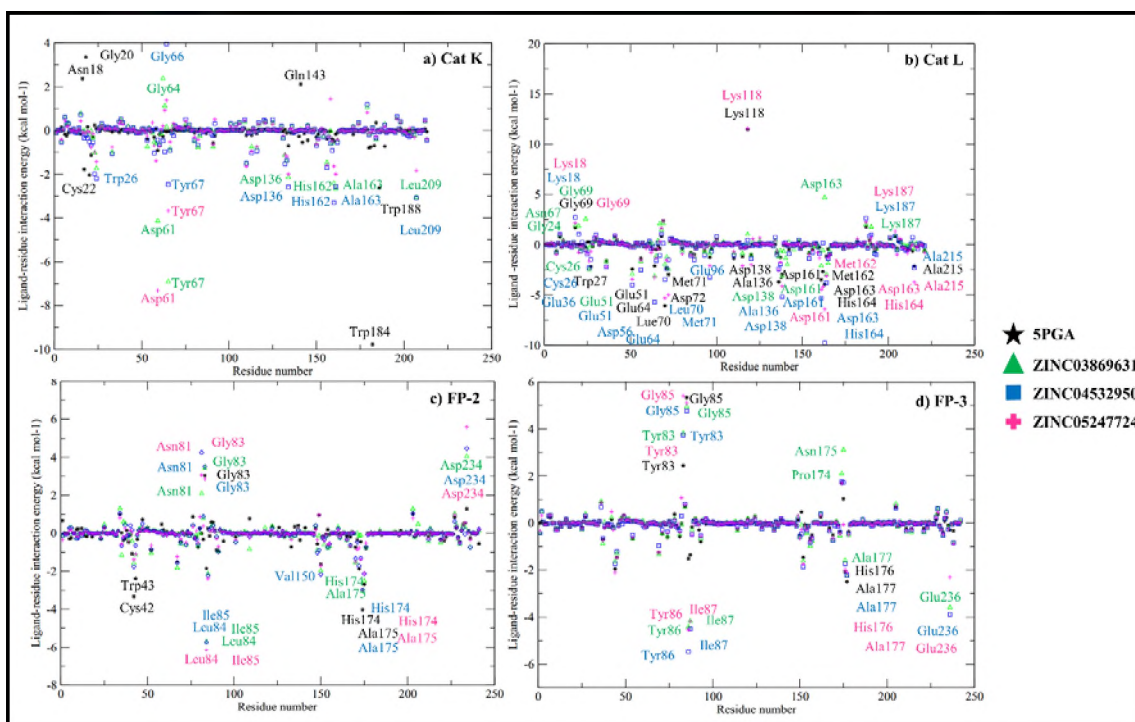


Figure 4.4: Per-residue decomposition analysis of 5PGA and the selected ZINC compounds when in complex with a) Cat K, b) Cat L, c) FP-2 and d) FP-3. Residues with values $> 0 \text{ kcal.mol}^{-1}$ impairs binding and vice versa.

To determine how the different subsites contributed to the BFE per ligand (Figure 2.6), net contribution of all residues per subsite was determined. Subsite S1, S2 and S3 exhibited

varying net energies (positive or negative) while S1' contributed to negative energy scores for the complexes of the three ligands and all proteases used in the study (Figure 4.5). As determined by the dynamical fingerprints obtained during MD simulations, majority of the S1' residues were directly involved in the formation of hydrogen or hydrophobic interactions with the ligands (Appendix 1K). In FP-2, S1 residues and S3 residues entirely contributed to a net positive energy as opposed to S2 and S1' for the three compounds. As for cathepsins, all the four subsites exhibited negative net energies in majority of the compounds an indication of poor selectivity. However, S2 and S3 residues in all non-human *plasmodial* proteases when in complex with CPG, CPH and CPI contributed to negative energies (Figure 4.5b and 4.5c). As seen in Figure 2.6, the composition of S2 is highly varied especially between human and *plasmodial* proteases. Thus by considering the observed contribution of the residues forming these subsites, essential structural and chemical information that should be present in ligands to achieve stronger interactions can be realised. . This is important in the rational drug design process of novel *plasmodial* cysteine protease inhibitors with increased selectivity towards the human proteases.

4.5.3 Structural features affecting BFE

Several ligand features that affected strength of interactions were observed. From the BFE results, the two sets of compounds (CPs and 5PGA-ZINC analogs) studied exhibited varying interaction energies mainly with the vdW and electrostatic energy terms. CPs exhibited lower vdW and electrostatic interactions compared to 5PGA and ZINC analogs thus stronger interactions. The vdW term was mainly influenced by size of the ligands which in turn affected the number of residues (Table 3.3 and Appendix K).

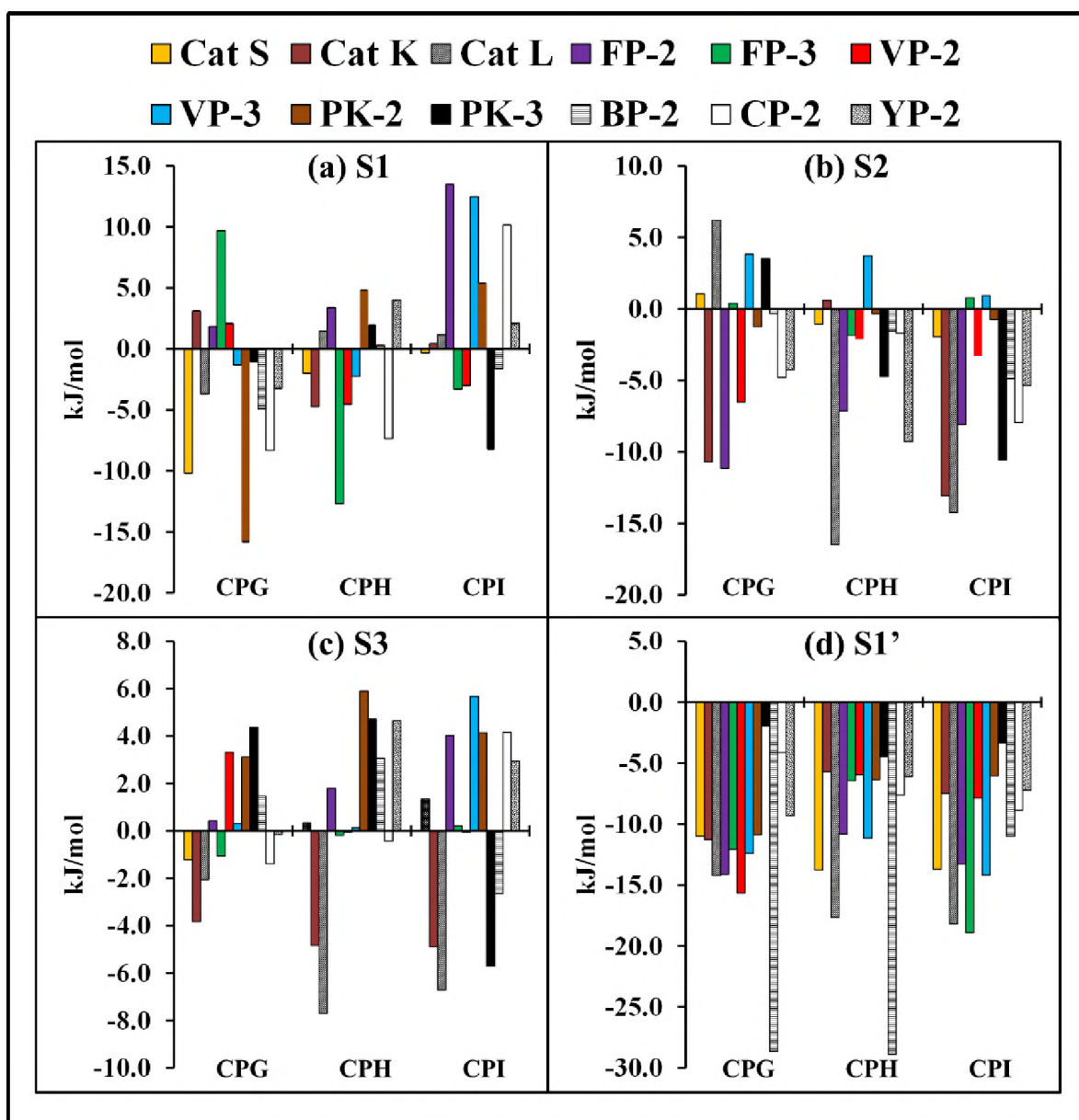


Figure 4.5: A detailed per-residue fingerprint showing the individual aa energy contributions to the binding of CPG (black), CPH (red) and CPI (green) with (a) Cat K (b) Cat L (c) FP-2 (d) FP-3 (e) VP-2 and (f) VP-3. Used in Musyoka TM *et al.*, 2015²²⁴.

For the electrostatic term, CPs exhibited lower values compared to 5PGA and its analogs. This was mainly determined by the number of hydrogen bonds formed between binding pocket residues and ligand atoms (Chapter 3 and Appendix 1K). CPs had the highest number of hydrogen bonds acceptors and donors thus contributing to higher electrostatic interactions.

4.6 Chapter conclusion

In line with the overall objective of this thesis, the current work aimed in characterizing the interactions between the various proteins and ligands so far studied. From the BFE results, a correlation between the chemical nature of the individual proteins studied (residue composition) and that of the ligands in association with their size was established. As the binding pockets of the proteins under study are mainly composed of hydrophobic residues, vdW interaction was identified as the main energetic term that promotes the binding of the ligands. In addition, the important value of hydrogen bonds (formed through hydrogen bond acceptors and donors) in enhancing the binding of the ligands was determined. An increased ability of a ligand to form hydrogen bonding with protein residues promoted binding. Thus the incorporation of hydrogen bonding centres in a ligand while designing is crucial in enhancing its potency. However, this should be carefully done in consideration to the acceptable numbers of such centres in a drug molecule (Table 3.1).

The subsequent decomposition of BFE to per residue contribution identified the key role of the different subsite residues. In all the proteins studied, S2 residues cumulatively contributed to a negative energy an indication of the significant role they play in enhancing ligand binding. An interesting feature observed in the case of S2 and S3 was the differential contribution by the S2 and S3 residues between human cathepsins (inhibited binding) and *plasmodial* proteases (promoted binding). This information is key and helpful in the CADD process of novel antimalarials against *plasmodial* proteases and with selectivity towards the human cathepsins. Although the binding free energies of the different complexes were successfully estimated via computational approaches, there is also a need to confirm these results via experimental techniques such as protein affinity studies or through more rigorous computational approaches such thermodynamic integration (TI).

CHAPTER 5

Docking studies: New *plasmodial* cysteine inhibitors from SANCDB

Over the last two decades, molecular docking has become one of the important strategies employed in the computer aided drug design (CADD) process mainly in the virtual screening stage. This is a computational approach used to determine how a putative ligand binds to an active pocket of a molecular target (protein). By this, the pose of a ligand (its orientation and conformational geometry) and binding affinity are determined via a search and scoring algorithm. Motivated by the results of a previous docking study using a set of 23 non-peptide compounds from South African natural sources where one potential hit was identified (5PGA), the current chapter aims to search for more similar hits from the South African Natural Compound Database (SANCDB). As this work is still ongoing, reported herein are the preliminary docking results of SANCDB compounds on four plasmodial proteins. These include the falcipains (FP-2 and FP-3) and knowlesipains (KP-2 and KP-3). Progressing studies will evaluate the activity of these compounds on the remaining plasmodial proteases as well as their selectivity on the human cathepsins. Additional studies to further characterize identified potential hits will be performed.

5.1 Introduction

To cut down on the huge costs involved in drug discovery, integrated efforts characterised by innovativeness and technology have been adopted in the pharmaceutical R&D³⁴². To meet the current economic pressure of availing drugs to the market, pharmaceutical companies in collaboration with academia have adopted a complex paradigm in the discovery of effective remedies to diseases. This involves the search of novel drug targets and new lead compounds³⁴³. One of the major pharmaceutical investments was the establishment of HTS technologies to identify new drug hits. This entails testing of large compound libraries for activity against selected biological targets using miniaturised assays coupled with large scale data analysis via automated platforms. Despite its substantial contributions in modern drug-discovery attempts, it has habitually failed in identification of potential leads^{344,345}.

An alternative and at times complimentary strategy to HTS that has also found its way in drug discovery process is use of virtual screening (VS) approach³⁴⁶ which is more information-rich compared to its predecessor. VS uses a combination of computational approaches to identify hits against specific targets. There are two main approaches used in VS *viz.* ligand based (LBVS) and structure based (SBVS). LBVS methods extrapolate from known active compounds which are used to search for structurally diverse compounds from chemical databases (known or in-house) with similar biological activity³⁴⁷. This is based on the similarity property principle by Johnson and Maggiora which postulates that similar structures have similar biological activity³⁴⁸. Through methods like fingerprint and pharmacophore based strategies, quantitative structure-activity relationship (QSAR), similarity searching and comparative molecular field analysis (CoMFA) large chemical databases can be screened leading to the identification of hits. Various similarity metrics and coefficients such as the Tanimoto coefficient³⁴⁹ are used to determine the likeness between

the query compound and the identified hit. A major characteristic of LBVS is that it does not require 3D structure of the specific molecular receptor.

For SBVS, it utilises information from the structure of a molecular target to identify hits which form energetically favourable interactions with the receptor's binding pocket residues³⁵⁰. This chapter will exclusively involve the SBVS approach to identify novel antimalarial hits. CADD has continuously adopted numerous computational modelling approaches to study the structure activity relationship (SAR) between molecules³⁵¹. Molecular docking is one of the most heavily used computational approach in SBVS whose main aim is to define the electrostatic and stereochemical attributes of a ligand within the constraints of the binding site of a receptor and to correctly approximate the binding affinity^{352,353}. For successful docking process, 3D structural information of the receptor must be available, a major draw backs to this approach. However, modern biomolecular spectroscopic (X-ray crystallography and NMR) and computational methodologies (homology modelling) are continuously providing a solution by availing reliable structures. Up to date, over 100,000 structures of macromolecular targets have been resolved through these methods providing vital structural information about key macromolecular drug targets. To identify lead compounds with high affinity on the drug targets, pharmaceutical companies analyse large compound libraries using docking approaches. This is important as it reduces the costs involved in chemical assays as all compounds without any desirable interactions profiles are dropped out. At the end of the process, only few compounds with high potentiality of becoming active leads are synthesized, assayed and if necessary optimized via chemical modification.

5.2 South African Natural Compound Database (SANCDB)

This is a growing fully referenced chemical database of compounds isolated from South African plant and marine sources²⁹. Currently, the database has 624 compounds (as of

January 2016) exhibiting a range of medicinal properties ranging from antimicrobial, anticancer and antidiabetic among others. So far, the database has shown to be an important resource in the search of hits against various drug targets ranging from human heat shock proteins (Hsp90)³⁵⁴, human cathepsins (unpublished work by an honours student), *Trypanosoma* Cat B-like proteins and human immunodeficiency virus proteases (unpublished work by a masters student). In the current work, which is still ongoing, SANCDB is used to mine for potential hits against *plasmodial* proteases.

5.3 Docking software

Up to date, over 60 docking programs and more than 30 scoring algorithms have been developed with varied accuracy and computational efficiency levels³⁵⁵. These can be classified into three major categories namely: those that utilise MD simulations and solvation models to determine absolute BFE (deterministic approaches), tools that use knowledge-based statistical potentials and those of empirical approaches based on regression approach. Most of these software treat the receptor as rigid and consider only the ligand flexibility which consist of translational and rotational degree of freedom. The most commonly used software tools include Automated Docking of Ligands to Macromolecules (AutoDock)³⁵⁶, DOCK³⁵⁷, Flexible Docking Method Using an Incremental Construction Algorithm (FlexX)³⁵⁸, Surflex-Dock³⁵⁹, Fast rigid exhaustive docking (FRED)³⁶⁰, Glide and Genetic Optimization for Ligand Docking (GOLD)³⁶¹. The original DOCK software first uses surface spheres to fill the binding pocket and their centers used to match atoms during rigid docking. Several subsequent modifications of DOCK have been implemented. These include DOCK 3.0 (molecular-mechanics force field scoring), DOCK 3.5 (energy minimization) and DOCK 4.0 which uses genetic algorithm (GA) leading to ligand conformational flexibility. In GOLD, both the ligand and the protein area around the binding site are set flexible. To determine the best docking result, the software evaluates the internal energy of the ligand,

sum of H-bonding and hydrophobic energies. A characteristic feature of all docking software established so far is that they consist of two key components; a search and a scoring algorithm. The search algorithm performs posing which includes the proper placement and positioning of a ligand within a putative binding site³⁵⁰. Commonly used docking tools determine the best structure by using simple scoring functions to sample the conformational space. This in turn affects the end result. To reduce these inaccuracies, classical physics has been incorporated in developing MD based algorithms. A good example of this category is CDOCKER which is based on the CHARMM force field.

5.4 AutoDock

Herein, AutoDock, an excellent and widely used open source (GNU General Public Licence and Apache Open Source License) program was explicitly used to perform docking studies. Since 1990 when it was first released, AutoDock has proven to be an effective tool and several other superior versions have been released the latest being AutoDock 4.2. It consists of three sub-tools namely AutoDock, AutoTors and AutoGrid. AutoTors which is the most simplest defines the rotatable bonds inherent in a given ligand. This in turn determines the degrees of freedom which determine the complexity of the simulations. Docking simulations involving ligands with ≤ 6 rotatable bonds are usually accurate and reasonably fast while those involving large ligands are prone to inaccuracies and computationally slow^{353,362}. Using AMBER force field and based on the macromolecular target, AutoGrid determines a 3D grid of interaction energy. AutoDock in turn performs the simulation by moving a ligand in any of the degrees of freedom followed by calculation of the new state's energy. It utilises a hybrid GA and Lamarckian genetic algorithm (LGA) to perform a local search at each new generation.

To determine the BFE, AutoDock scoring function uses a rather inexpensive force field consisting of both semi-empirical and molecular mechanics terms. To evaluate binding, the scoring function estimates the intramolecular energetics involved in the transition from the unbound to bound conformation of the protein (P) and ligand (L). The force field includes the conformational entropy lost during binding (ΔS_{conf}) six pair-wise energetic terms (V) as shown in equation 5.1. Each term includes evaluation for desolvation, electrostatics, hydrogen bonding, dispersion and repulsion.

$$\Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{P-P}) + (V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}) \quad (5.1)$$

A recent successor of AutoDock is Vina which has a knowledge-based statistical scoring function which improves its prediction speed. This is by utilizing a simplified scoring function and multi-threading in cases where multiple computer cores are available³⁶³.

5.5 Docking types

There are three main docking approaches namely rigid and flexible ligand and flexible docking³⁶⁴. In rigid docking, both the internal geometry of the protein and the ligand are kept fixed. It is based on the lock and key theorem proposed by Fischer in 1890. In flexible ligand docking the ligand is kept flexible and the energies from different conformations determined. So far, these two approaches are the most commonly used. In flexible docking, both the receptor and the ligands flexibility are considered. Developing algorithms that can effectively determine the flexibility of a ligand and a receptor has remained a big challenge owing to the expensive nature of such simulations. To emulate receptor flexibility, an ensemble of static structures of a given receptor with different conformations are experimentally determined and then used in the docking process³⁶⁵. An alternative to this is to generate rotamer libraries of binding pocket residues side chains and search for energetically accurate protein conformation³⁶⁶.

5.6 Methodology

Figure 5.1 shows a workflow of the approaches used in this study.

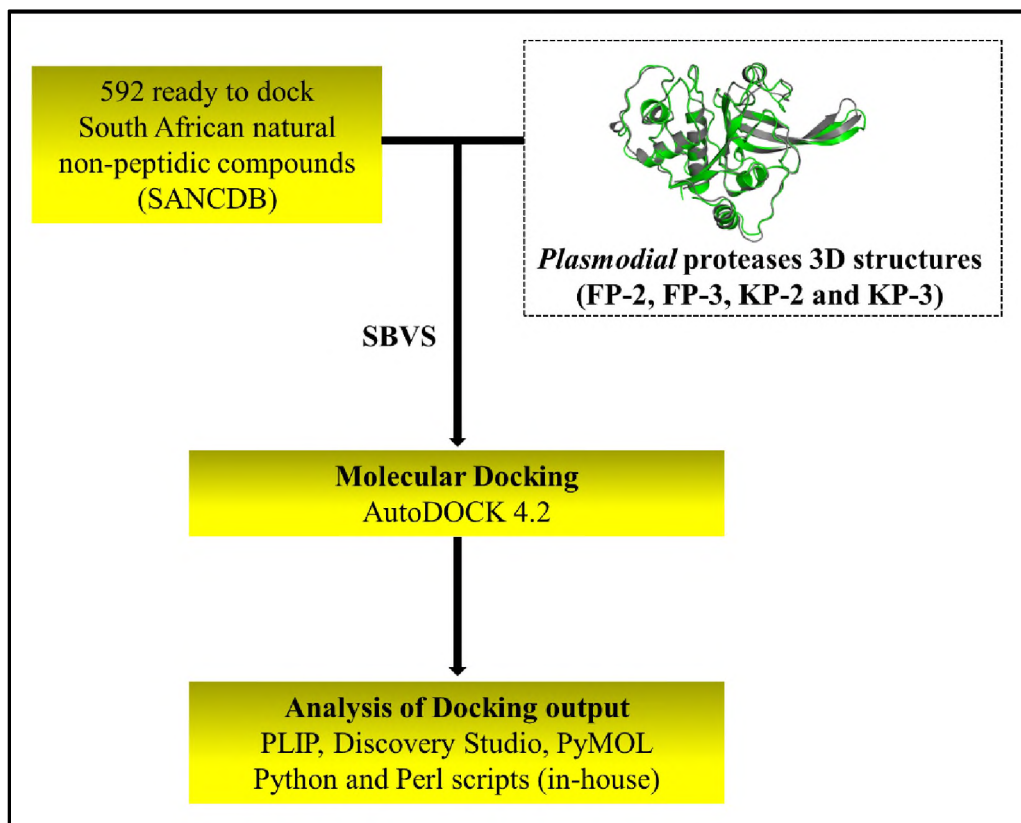


Figure 5.1: A diagrammatic representation of the different steps and tools used for docking studies. Adapted from Musyoka *et al.*, 2016³⁰⁰.

5.6.1 Protein structure data and ligands

Crystallographic structure files for FPs (FP-2 [2OUL] and FP-3 [3BWK]) were retrieved from the Protein Data Bank (PDB). High quality homology models of KPs (KP-2 and KP-3) were previously calculated using MODELLER version 9.10 as described by Musyoka *et al.*, 2015²²⁴. Prior to docking, all crystallographic water molecules and bound ligands were removed on all 3D structures obtained from PDB. A set of 592 compounds (some compounds could not be minimized due to their huge size) were retrieved from SANCDDB in the PDB format. These compounds were ready to dock as they had been minimised previously using the General Atomic and Molecular Electronic Structure System (GAMESS)³⁶⁷.

5.6.2 Docking studies

Docking simulations using AutoDock 4.2 consists of three major steps namely; preparation of protein and ligand files, determination of affinity maps via a 3D grid and docking simulation under predefined parameters. An important preparatory process known as docking validation is crucial in order to ascertain if the docking parameters adopted are accurate. This involves extracting and redocking a ligand co-crystallised with a 3D structure of the protein being studied. However, in the current work, this process had been accomplished previously²¹⁵.

5.6.2.1 Preparation of protein and ligand files

Using AutoDock Python scripts (*prepare_receptor4.py* and *prepare_ligand4.py*), polar hydrogens were first added to all ligand (L.pdb) and protein (P.pdb) files generating corresponding LH.pdb and PH.pdb files. Subsequently, charge (Gasteiger-Hückel) and atom type information were added coupled with the merging of all non-polar hydrogens for calculation of affinity maps. The hydrogenised pdb files were finally converted to rigid conformations known as pdbqt files. For the ligands, torsions around the rotatable bonds were automatically set during the conversion process³⁵⁶.

5.6.2.2 Grid evaluation and affinity maps determinations

Based on the each protein coordinates, AutoGrid 4.2 was used to determine 3D grid of interaction energies using AMBER force field. Interaction energies between the probe and surrounding residues at each grid point were determined and stored in a table. The grid box dimensions were set as 70, 70, 65 (Å) along the x, y and z directions and a spacing of 0.3472 Å. This ensured that each atom types from the ligands and those for dispersion/repulsion and electrostatic interactions were chosen sufficiently large enough not just to cover the active site but also important surrounding areas. Cys42 of FP-2 (131.759, 83.811, -180.551) and corresponding positions in FP-3 (6.856, -20.322, 44.439), KP-2 (5.225, -18.497, 36.209) and KP-3 (-0.498, -12.204, 32.183) were chosen to be the centroid point of the grid boxes. The

grid box spanned an area of residues around a 12Å radius. Using an *ad hoc* Python script, the process was fully automated.

5.6.3 Docking simulation

For each ligand and protein, a docking parameter file (DPF) was generated using an AutoDock Python script (*autodpf.py*). The parameters selected for docking were: GA was used for conformational space search while for protein-ligand conformational search, LGA was utilised. The population size was set at 150, 100 GA runs, maximum energy evaluations of 450,000 and maximum number of generations set at 27,000. Cluster analysis for docked results was done using a root mean square (RMS) tolerance of 2.0 Å.

5.6.4 Analysis

Using AutoDock Python scripts namely *autodlg_analyzer.py* and *write_lowest_energy_ligand.py* were used to determine the best conformation and corresponding estimated interaction energy. A summary file containing the energy score of each pose calculated was generated for each ligand using a AutoDock Python script, *summarize_docking.py*. Using an *ad hoc* Python script, all docking files (dlg) were parsed in an automated manner to extract the conformation of each ligand with lowest energy for each protein. The output was parsed on to another python script (*write_lowest_energy_ligand.py*) to determine the ligand with lowest energy overall. Also determined by AutoDock is the inhibition constant (K_i) which is $\exp(\Delta G/(R \cdot T))$ where T is temperature and R is gas constant. Best ligand conformation was converted to PDB format using the Python script *pdbqt_to_pdb.py* and Discovery Studio version 4.1 (Accelrys Software Inc. San Diego) and protein-ligand interaction profiler (PLIP)³⁶⁸ were used to determine the type of interactions such as hydrogen bonds, hydrophobic as well as π - π interactions and residues involved. For visualization, PyMOL was utilised.

5.6.5 System specifications

All docking experiments were performed on an in house Linux cluster pre-installed with the required AutoDock 4.2 software.

5.7 Results and Discussion

5.7.1 Identification of best hits

Table 5.1: Best hits against *plasmodial* cysteine proteases identified from SANCDB with interaction energy of < -10.0 kcal/mol

Protein	Compound SANCDB ID	Interaction energy (kcal/mol)
FP-2	SANC00686	-10.12
	SANC00220	-10.23
	SANC00518*	-10.24
	SANC00454	-10.52
FP-3	SANC00388	-10.03
	SANC00478	-10.11
	SANC00491	-10.11
	SANC00288	-10.27
	SANC00511	-10.44
	SANC00289	-10.45
	SANC00287	-10.46
	SANC00290	-10.58
	SANC00518*	-10.60
KP-2	SANC00518*	-10.02
	SANC00421	-10.08
	SANC00616	-10.35
KP-3	SANC00511	-10.07
	SANC00480	-10.35
	SANC00512	-10.46
	SANC00389	-10.79
	SANC00518*	-10.95

**Compound with common activity. Highlighted are the best hits and corresponding*

To screen for antimalarial activity, all SANCDB compounds (at that time) were docked against four *plasmodial* cysteine proteases. It was not a prerequisite that the natural compounds used had antimalarial activity tested before. Based on the docking energies,

several potential hits with interesting inhibitory profiles on *plasmodial* proteases tested were identified. Table 5.1 shows hits with interaction energies of < -10.00 kcal/mol against FP-2, FP-3, KP-2 and KP-3. Interestingly, compound SANC00518 had high *in silico* activity against all the four proteases utilised for screening. For FP-2, the best hit was compound SANC00454 with interaction energy of -10.52 kcal/mol while in KP-2 it was SANC00616 with interaction energy of -10.35 kcal/mol. SANC00518 was the best hit for both FP-3 and KP-3 with interaction energies of -10.60 and -10.95 kcal/mol respectively.

5.7.2 Molecular interactions of best hits

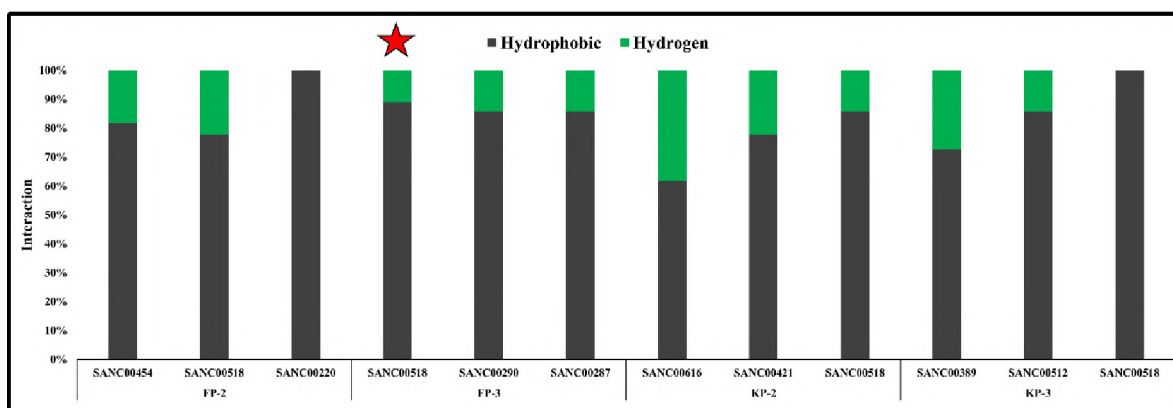


Figure 5.2: A stacked column chart showing the type and percentage of residues interacting with the best three hits identified per protein as determined by protein-ligand interaction profiler (PLIP) analysis software. Indicated with star is ligand exhibiting π - π interaction.

All the identified hits fitted perfectly on the “trench” like active site of the proteases studied thus the excellent interaction energies. Majority of the interactions formed between the protein residues and ligand atoms were mainly hydrophobic in nature (Figure 5.2). As seen in Table 5.1, the identified hits bound to the different proteins with the same strength as defined by the interaction energies. A comparison of the best three hits per protein and their resulting interactions with corresponding protein residue shows that the quantity of the main interactions (hydrogen and hydrophobic) were comparable. In FP-2 (Figure 5.3A & B), SANC00454 mainly formed hydrophobic bonds with S2 residues (L84, I85, L172, A175 and D234). This is an interesting feature as it has been shown that S2 residues besides being key

players in the ligand binding process, they also confer selectivity²²⁴ as seen in Chapter 4. One terminal end of SANC00454 interacted with the deepest residue of the binding pocket (D234). Other major hydrophobic interactions were formed with S1' residues (V152 and W206). In addition to hydrophobic interactions, two hydrogen bonds were formed with residues I85 (S2) and W206 (S1'). Cumulatively, these interactions resulted to an inhibition constant of nanomolar range (19.46 nM). In FP-3 (Figure 5.3C & D), SANC00518 interacted with all subsite residues except S3 via hydrophobic interactions [Q38 (S1), I87, A177 (S2), H176 and W208 (S1')]. H176 in addition formed hydrogen bonding and π -stack interactions with the ligand. Estimated inhibition constant of this ligand was also up to the nanomolar level (17.04 nM). SANC00616 and KP-2 exhibited the highest number of the hydrophobic interactions and hydrogen bonds (Figure 5.4A and B). The subsite interacting residues were Q37, D82 (S1), L85, I86, P173, N174, A176, E235 (S2), and W207 (S1'). In addition, key subsite residues including Q37, C40, D82 (S1), I86, E235 (S2) and H175, W207 (S1') were involved in the formation of hydrogen bonding with the ligand. The estimated inhibition constant was 25.72 nM. For KP-3, SANC00518 never exhibited hydrophobic interactions with S1 and S3 subsites residues which are known to impair binding thus the low observed docking inhibition of 9.37 nM (Figure 5.4C and D)²²⁴. It mainly interacted with S2 residues (F83, I84, N148, T171, N172 and A174). Although docking studies using other proteins is yet to be performed, it will be interesting to find how the identified ligand would interact with corresponding S2 residues especially in human Cat K and L (Chapter 2).

The identified hits had no previous record of antimalarial activity identified. For SANC00454, whose name is Pregn-5-en-20-one and closely related to 5 α -Pregna-1,20-dien-3-one (5PGA [SANC00146]), previously shown to possess potential activity against *plasmodial* proteases (Chapter 4). They both belong to the pregnadiene sterol group of compounds although they were isolated from different source organisms; *Capnella*

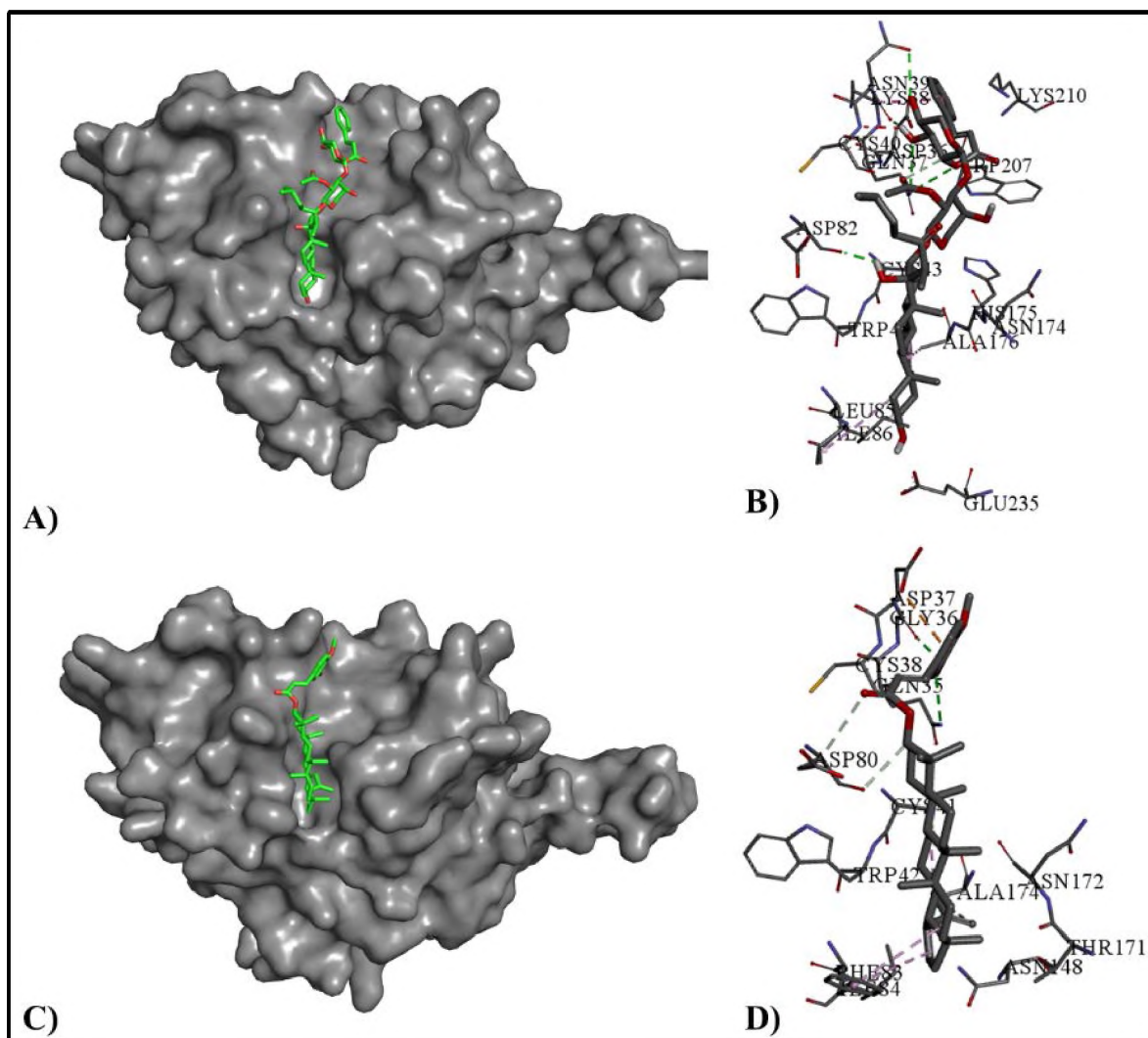


Figure 5.4: A surface presentation and corresponding 2D interaction map of KP-2 (A and B) and KP-3 (C and D) when docked with SANC00616 and SANC00518 respectively. Dotted green lines depict hydrogen bonding while grey hydrophobic interactions.

5.7.3 Comparison with CPs, 5PGA and ZINC hits

A comparison of current docking results with those of the previously studied compounds (CPs, 5PGA and its analogs), the following can be established; 1) the best hits from CPs (CPG, CPH and CPI) have similar interaction energies as those of the newly identified hits between corresponding proteins. 2) the size and the chemical groups present in a given ligand are the key factors affecting its potency. 3) the compounds identified in this chapter present themselves as better alternatives to 5PGA and its ZINC analogs which require further assessment on their suitability as novel antimalarial hits.

Both CPs and the newly identified had an extended structure which enabled them to interact with majority of the subsite residues leading to stronger interactions (Figure 3.5). This was in comparison to 5PGA and its analogs. In terms of aa interaction network, the compounds interacted with same residues as with the CPs. It will be important to investigate if the same residues contribute to the BFE in a similar manner when interacting with the new set of ligands (Figure 3.24 and Chapter 4).

5.8 Chapter conclusion

Although the current work reported herein is still ongoing, interesting hits with better inhibitory potencies against *plasmodial* proteases compared to those of 5PGA and its analogs have already been identified. Thus, the present results indicate novel compounds with inhibitory activity against *plasmodial* cysteine proteases can be mined from SANCDB. This underscores the importance of SANCDB as a tool in the drug discovery process. To determine if identified hits can target homologs from other *Plasmodium* spp., the docking process is currently being extended to VP-2, VP-3, CP-2, BP-2 and YP-2. In addition, it is also necessary to determine the inhibitory profiles of selected hits against Cat K and Cat L. This is essential to evaluate their selectivity towards host's proteases machinery. Additional wet laboratory assays involving collaborations with other research groups may be undertaken in the near future to confirm the findings. Where necessary chemical modifications to enhance hit inhibitory activity and selectivity will be performed.

CHAPTER 6

Conclusions and future prospects

To determine the potentiality of plasmodium cysteine proteases as drug targets and try to identify novel non-peptide inhibitors using in silico approaches. Through various in silico approaches namely docking, sequence and structural analysis, MD simulation and BFE studies key insight to CADD process of new inhibitors against the proteases were identified. This chapter summarises the key findings identified. Also, future prospects utilizing both in silico and in vivo approaches for the validation of the current findings and attainment of newer hits will be addressed.

6.1 Conclusions

For the first time, the present work presents a detailed *in silico* approach towards the discovery of novel protease inhibitors targeting malarial proteins from different *Plasmodium* species. Although FP-2 and FP-3 have been considered as attractive targets for antimalarial drug development, there has never been a drug developed against them despite numerous attempts. Previous attempts mainly resulted in peptidic compounds^{184,205–207}. However, due to their inherent chemical and pharmacological profiles, the results were futile as the compounds were prone to degradation by host enzymes and were excessively large.

To overcome these challenges, the search for non-peptidic compounds with inhibitory potency against *plasmodial* cysteine proteases is gaining momentum^{371,372}. However, these studies mainly target FP-2 and or FP-3 thus not solving the problem at hand exclusively. In addition, the use of dockings solely for drug discovery and development is not adequate enough as proteins are dynamic structures. The current work being first of its kind goes beyond these limitations and successfully introduces the aspect of MD simulations, determination of energy of intermolecular associations. In addition, another unique feature of this study is the targeting of several other *plasmodial* proteases besides FP-2 and FP-3. The role of such complex computational experiments (MD simulations and BFE calculations) cannot be ignored.

As FP-2 and FP-3 are close homologs to human cathepsins and that they both share more or less catalytic mechanism, a detailed analysis involving structural information, physiochemical properties, phylogenetic and motif elucidation was performed. This was to determine any vital difference that would be useful in the attainment of drug selectivity. From the results (Chapter 2); there exists sequence variation between *plasmodial* proteases and human cathepsins. This can be confirmed by the SI values (Table 2.1), MSA information (Figure 2.6) and phylogenetic output (Figure 2.8). In addition, differences between individual

sequences of either groups (*plasmodial* or human cathepsins) were also observed which can be clearly seen in the dendrogram (Figure 2.8). Differential location and distribution of motifs was also observed, with human cathepsins and *plasmodial* proteases having two and three unique motifs respectively. The distinguishing motifs were located in the R-domain of the protein structures. The functional significance of this is not clear and further analysis is needed to ascertain if this could be utilised in attaining drug selectivity. Of the physicochemical properties studied, aromaticity, GRAVY were fairly similar while significant differences were observed with the molecular weight and pI properties (Table 2.2). For molecular weight, the observed difference was expected as *plasmodial* sequences are longer (possess two inserts) compared to cathepsins. In terms of pI, there was no clear difference or trend between the two groups. Further investigation for the wide variation of *plasmodial* protease pI is necessary. However, in the case of cathepsins, Cat L which is the main lysosomal cathepsins had pI that matched its native environment (pI=4.64), while Cat K and S which are mainly localized on other body cells such as osteoclasts had high pI values (8.92 and 7.64 respectively).

Despite the observed differences, there was little or no selectivity observed when the proteinligand complexes were studied via MD simulations and BFE calculations. However, previous pre-clinical studies involving cysteine protease inhibitors targeting *trypanosomal* and *plasmodial* parasites and with little or no selectivity were well tolerated in animal models. This was because cathepsins are present in higher concentrations compared to that from the parasite and the redundant nature of mammalian cysteine proteases family inhibiting the roles of one cathepsin, its roles will be played by another³⁷³. The findings of this study could be extended to the generation of novel drugs against cathepsins as they have been found to play roles in antigen presentation, bone resorption and pro-hormone activation.

These processes have been found to play a role in the progression of a variety of disease states such as rheumatoid arthritis, osteoporosis and autoimmune maladies.

Using a different approach to mine for FP-2 inhibitors, Cátia *et al.*, performed a 3D-QSAR on peptidyl vinyl sulfone derivatives and determined the major structural requirements necessary for optimal activity of ligands binding its pocket³⁷⁴. A similar approach by Wang *et al.*, using heteroarylnitrile derivatives made similar observations where different subsites preferred chemical groups with certain properties²⁰⁹. A great consistency exists between the previous findings and our current results from docking, MD and BFE analysis. In the per-residue energy decomposition of F2-CPI, a total of 10 residues including Asp35, Cys42, Leu84, Val152, Ser153, Asp154, Asp155, Asn173, His174 and Ala175 were identified as key to the binding process as observed with the 3D-QSAR approach.

6.2 Future prospects

In order to confirm the reliability of the identified hits identified through the current *in silico* approaches (molecular docking, MD simulations and BFE studies), incorporation of wet laboratory assays involving whole parasites and recombinant proteases is necessary. This will be performed through collaborative initiatives with research groups working with the listed models.

To improve their drug-likeness, hit optimization through additional chemical modifications are necessary. In the case of CPG, CPH and CPI their molecular weights need to be lowered to the acceptable values. For 5PGA and its ZINC analogs, the elongation of their main chains and introduction of additional side groups to increase their molecular size while increasing the number of hydrogen bond donors and hydrogen bond acceptors are important. This should be done carefully whilst lowering their high LogP values.

To facilitate screening of large chemical databases for potential antimalarial hits, the MD-pipeline needs to be equipped with BFE evaluation part and a complete analysis tools leading to a complete standalone MD-BFE analysis tool. This is ongoing process currently.

The docking of SANCDB compound dataset will be performed to the other proteins and subsequently perform MD and BFE calculations.

References

1. WHO, World Malaria Report 2013. (2013) Available at: http://www.who.int/malaria/publications/world_malaria_report_2013/wmr2013_no_pr_ofiles.pdf?ua=1. (Accessed: 26th July 2015)
2. Hall, N. Genomic insights into the other malaria. *Nat. Genet.* **44**, 962–963 (2012).
3. Ashley, E., McGready, R., Proux, S. & Nosten, F. Malaria. *Travel Med. Infect. Dis.* **4**, 159–173 (2006).
4. Prugnolle, F. *et al.* A fresh look at the origin of *Plasmodium falciparum*, the most malignant malaria agent. *PLoS Pathog.* **7**, e1001283 (2011).
5. Cox-Singh, J. & Singh, B. Knowlesi malaria: newly emergent and of public health importance? *Trends Parasitol.* **24**, 406–410 (2008).
6. White, N. J. *Plasmodium knowlesi*: the fifth human malaria parasite. *Clin. Infect. Dis.* **46**, 172–173 (2008).
7. Sabbatani, S., Fiorino, S. & Manfredi, R. The emerging of the fifth malaria parasite (*Plasmodium knowlesi*): a public health concern? *Brazilian J. Infect. Dis.* **14**, 299–309 (2010).
8. Cox-Singh, J. *et al.* *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. *Clin. Infect. Dis.* **46**, 165–71 (2008).
9. Greenwood, B. M. *et al.* Malaria: progress, perils, and prospects for eradication. *J. Clin. Invest.* **118**, 1266–1276 (2008).
10. Langhorne, J. *et al.* The relevance of non-human primate and rodent malaria models for humans. *Malar. J.* **10**, 23 (2011).
11. Winzeler, E. A. Malaria research in the post-genomic era. *Nature* **455**, 751–6 (2008).
12. Wirth, D. F. Biological revelations. *Nature* **419**, 495–496 (2002).
13. Schuster, F. L. Cultivation of *plasmodium* spp. *Clin. Microbiol. Rev.* **15**, 355–364 (2002).
14. Champagne, D. E. Antihemostatic molecules from saliva of blood-feeding arthropods. *Pathophysiol. Haemost. Thromb.* **34**, 221–227 (2006).
15. Mota, M. M. & Rodriguez, A. Migration through host cells: the first steps of *Plasmodium* sporozoites in the mammalian host. *Cell. Microbiol.* **6**, 1113–1118 (2004).
16. Markus, M. B. Malaria: origin of the term ‘hypnozoite’. *J. Hist. Biol.* **44**, 781–6 (2011).
17. Kariuki, M. M., Li, X., Yamodo, I., Chishti, A. H. & Oh, S. S. Two *Plasmodium falciparum* merozoite proteins binding to erythrocyte band 3 form a direct complex. *Biochem. Biophys. Res. Commun.* **338**, 1690–5 (2005).
18. Kadekoppala, M. & Holder, A. A. Merozoite surface proteins of the malaria parasite: the MSP1 complex and the MSP7 family. *Int. J. Parasitol.* **40**, 1155–61 (2010).
19. Chou, A. C. & Fitch, C. D. Mechanism of hemolysis induced by ferriprotoporphyrin

- IX. *J. Clin. Invest.* **68**, 672–677 (1981).
20. Klouche, K. *et al.* Mechanism of in vitro heme-induced LDL oxidation: effects of antioxidants. *Eur. J. Clin. Invest.* **34**, 619–25 (2004).
 21. Lin, J. -w. *et al.* Replication of Plasmodium in reticulocytes can occur without hemozoin formation, resulting in chloroquine resistance. *J. Exp. Med.* **212**, 893–903 (2015).
 22. Salmon, B. L., Oksman, A. & Goldberg, D. E. Malaria parasite exit from the host erythrocyte: a two-step process requiring extraerythrocytic proteolysis. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 271–276 (2001).
 23. Adams, S., Brown, H. & Turner, G. Breaking down the blood–brain barrier: signaling a path to cerebral malaria? *Trends Parasitol.* **18**, 360–366 (2002).
 24. Francis, S. E., Sullivan Jr, D. J. & Goldberg, D. E. Hemoglobin metabolism in the malaria parasite Plasmodium falciparum. *Annu. Rev. Microbiol.* **51**, 97–123 (1997).
 25. CDC, Malaria. (2015) Available at: <http://www.cdc.gov/malaria/about/disease.html>. (Accessed 10th October 2015)
 26. Butler, M. S., Robertson, A. A. B. & Cooper, M. A. Natural product and natural product derived drugs in clinical trials. *Nat. Prod. Rep.* **31**, 1612–61 (2014).
 27. Koehn, F. E. & Carter, G. T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.* **4**, 206–20 (2005).
 28. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14**, 111–129 (2015).
 29. Hatherley, R. *et al.* SANCDB: a South African natural compound database. *J. Cheminform.* **7**, 29 (2015).
 30. Wells, T. N. C. Natural products as starting points for future anti-malarial therapies: going back to our roots? *Malar. J.* **10 Suppl 1**, S3 (2011).
 31. Schlitzer, M. Malaria Chemotherapeutics Part I: History of Antimalarial Drug Development, Currently Used Therapeutics, and Drugs in Clinical Development. *ChemMedChem* **2**, 944–986 (2007).
 32. Cousins, K. R. ChemDraw Ultra 9.0. CambridgeSoft, 100 CambridgePark Drive, Cambridge, MA 02140. www.cambridgesoft.com. See Web site for pricing options. *J. Am. Chem. Soc.* **127**, 4115–4116 (2005).
 33. Foley, M. Quinoline Antimalarials Mechanisms of Action and Resistance and Prospects for New Agents. *Pharmacol. Ther.* **79**, 55–87 (1998).
 34. Cox, F. E. History of the discovery of the malaria parasites and their vectors. *Parasit. Vectors* **3**, 5 (2010).
 35. Cox, F. E. History of the discovery of the malaria parasites and their vectors. *Parasit. Vectors* **3**, 5 (2010).
 36. Vennerstrom, J. L., Makler, M. T., Angerhofer, C. K. & Williams, J. A. Antimalarial dyes revisited: xanthenes, azines, oxazines, and thiazines. *Antimicrob. Agents Chemother.* **39**, 2671–7 (1995).
 37. Greenwood, D. Conflicts of interest: the genesis of synthetic antimalarial agents in

- peace and war. *J. Antimicrob. Chemother.* **36**, 857–72 (1995).
38. Coatney, G. R. Pitfalls in a discovery: the chronicle of chloroquine. *Am. J. Trop. Med. Hyg.* **12**, 121–8 (1963).
 39. Taylor, W. R. J. & White, N. J. Antimalarial drug toxicity: a review. *Drug Saf.* **27**, 25–61 (2004).
 40. da Silva, R. & Hochman, G. [A method called Pinotti: medicated salt, malaria, and international health (1952-1960)]. *História, ciências, saúde--Manguinhos* **18**, 519–43 (2011).
 41. Payne, D. Spread of chloroquine resistance in *Plasmodium falciparum*. *Parasitol. Today* **3**, 241–246 (1987).
 42. Tu, Y. The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nat. Med.* **17**, 1217–20 (2011).
 43. Liao, F. Discovery of Artemisinin (Qinghaosu). *Molecules* **14**, 5362–5366 (2009).
 44. Klayman, D. L. *et al.* Isolation of Artemisinin (Qinghaosu) from *Artemisia annua* Growing in the United States. *J. Nat. Prod.* **47**, 715–717 (1984).
 45. Warhurst, D. C. New drugs and their potential use against drug-resistant malaria. *Ann. Ist. Super. Sanita* **21**, 327–36 (1985).
 46. Maude, R. J., Woodrow, C. J. & White, L. J. Artemisinin antimalarials: preserving the "magic bullet". *Drug Dev. Res.* **71**, 12-19 (2009).
 47. Medhi, B., Patyar, S., Rao, R. S., Byrav D S, P. & Prakash, A. Pharmacokinetic and toxicological profile of artemisinin compounds: an update. *Pharmacology* **84**, 323–32 (2009).
 48. WHO, Q&A on artemisinin resistance. (2015) Available at: http://who.int/malaria/media/artemisinin_resistance_qa/en/. (Accessed: 8th August 2015)
 49. WHO, Emergence and spread of artemisinin resistance calls for intensified efforts to withdraw oral artemisinin-based monotherapy from the market. (2014) Available at <http://www.who.int/malaria/publications/atoz/policy-brief-withdrawal-of-oral-artemisinin-based-monotherapies/en/>. Accessed: 10th October 2015)
 50. WHO, Overview of malaria treatment. (2015) Available at <http://www.who.int/malaria/areas/treatment/overview/en/>. Accessed: 23rd November 2015
 51. Schaefer, B. *Natural Products in the Chemical Industry*. (Springer, 2015).
 52. Loub, W. D., Farnsworth, N. R., Soejarto, D. D. & Quinn, M. L. NAPRALERT: computer handling of natural product research data. *J. Chem. Inf. Model.* **25**, 99–103 (1985).
 53. Chen, X. *et al.* Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. *Br. J. Pharmacol.* **149**, 1092–103 (2006).
 54. Valli, M. *et al.* Development of a Natural Products Database from the Biodiversity of Brazil. *J. Nat. Prod.* **76**, 439–444 (2013).
 55. Ntie-Kang, F. *et al.* ConMedNP: a natural product library from Central African

- medicinal plants for drug discovery. *RSC Adv.* **4**, 409–419 (2014).
56. Ntie-Kang, F. *et al.* AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLoS One* **8**, e78085 (2013).
 57. Saxena, S., Pant, N., Jain, D. C. & Bhakuni, R. S. Antimalarial agents from plant sources. *Curr. Sci.* **85**, 1314–1329 (2003).
 58. Frederich, M., Tits, M. & Angenot, L. Potential antimalarial activity of indole alkaloids. *Trans. R. Soc. Trop. Med. Hyg.* **102**, 11–19 (2008).
 59. Kaur, K., Jain, M., Kaur, T. & Jain, R. Antimalarials from nature. *Bioorg. Med. Chem.* **17**, 3229–56 (2009).
 60. Novoa, E. M. *et al.* Analogs of natural aminoacyl-tRNA synthetase inhibitors clear malaria in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5508–17 (2014).
 61. Phyto, A. P. *et al.* Antimalarial activity of artefenomel (OZ439), a novel synthetic antimalarial endoperoxide, in patients with *Plasmodium falciparum* and *Plasmodium vivax* malaria: an open-label phase 2 trial. *Lancet Infect. Dis.* (2015). doi:10.1016/S1473-3099(15)00320-5
 62. BRUCE-CHWATT, L. J. Classification of antimalarial drugs in relation to different stages in the life-cycle of the parasite: commentary on a diagram. *Bull. World Health Organ.* **27**, 287–90 (1962).
 63. Antimalarial Drugs - Malaria Site. at <<http://www.malariasite.com/malaria-drugs/>>
 64. Schmid-Hempel, P. Immune defence, parasite evasion strategies and their relevance for ‘macroscopic phenomena’ such as virulence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364**, 85–98 (2009).
 65. Leirião, P., Rodrigues, C. D., Albuquerque, S. S. & Mota, M. M. Survival of protozoan intracellular parasites in host cells. *EMBO Rep.* **5**, 1142–7 (2004).
 66. Severini, C. & Menegon, M. Resistance to antimalarial drugs: An endless world war against *Plasmodium* that we risk losing. *J. Glob. Antimicrob. Resist.* **3**, 58–63 (2015).
 67. Travassos, M. A. & Laufer, M. K. Resistance to antimalarial drugs: molecular, pharmacologic, and clinical considerations. *Pediatr. Res.* **65**, 64R–70R (2009).
 68. Fidock, D. A. *et al.* Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* **6**, 861–71 (2000).
 69. Cooper, R. A. Alternative Mutations at Position 76 of the Vacuolar Transmembrane Protein PfCRT Are Associated with Chloroquine Resistance and Unique Stereospecific Quinine and Quinidine Responses in *Plasmodium falciparum*. *Mol. Pharmacol.* **61**, 35–42 (2002).
 70. Sidhu, A. B. S., Valderramos, S. G. & Fidock, D. A. *pfmdr1* mutations contribute to quinine resistance and enhance mefloquine and artemisinin sensitivity in *Plasmodium falciparum*. *Mol. Microbiol.* **57**, 913–26 (2005).
 71. Nkrumah, L. J. *et al.* Probing the multifactorial basis of *Plasmodium falciparum* quinine resistance: Evidence for a strain-specific contribution of the sodium-proton exchanger PfNHE. *Mol. Biochem. Parasitol.* **165**, 122–131 (2009).

72. Fisher, N. *et al.* Cytochrome b mutation Y268S conferring atovaquone resistance phenotype in malaria parasite results in reduced parasite bc1 catalytic turnover and protein expression. *J. Biol. Chem.* **287**, 9731–41 (2012).
73. Happi, C. T. *et al.* Association between mutations in plasmodium falciparum chloroquine resistance transporter and p. falciparum multidrug resistance 1 genes and in vivo amodiaquine resistance in p. falciparum malaria-infected children in nigeria. *Am J Trop Med Hyg* **75**, 155–161 (2006).
74. Triglia, T. & Cowman, A. F. The mechanism of resistance to sulfa drugs in Plasmodium falciparum. *Drug Resist. Updat.* **2**, 15–19 (1999).
75. Sisowath, C. *et al.* The role of pfmdr1 in Plasmodium falciparum tolerance to artemether-lumefantrine in Africa. *Trop. Med. Int. Heal.* **12**, 736–742 (2007).
76. Eastman, R. T., Dharia, N. V, Winzeler, E. A. & Fidock, D. A. Piperaquine resistance is associated with a copy number variation on chromosome 5 in drug-pressured Plasmodium falciparum parasites. *Antimicrob. Agents Chemother.* **55**, 3908–16 (2011).
77. Nzila, A. The past, present and future of antifolates in the treatment of Plasmodium falciparum infection. *J. Antimicrob. Chemother.* **57**, 1043–54 (2006).
78. Mbengue, A. *et al.* A molecular mechanism of artemisinin resistance in Plasmodium falciparum malaria. *Nature* **520**, 683–687 (2015).
79. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
80. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498–511 (2002).
81. Holt, R. A. *et al.* The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**, 129–49 (2002).
82. Shukla, S. K., Murali, N. S. & Brilliant, M. H. Personalized medicine going precise: from genomics to microbiomics. *Trends Mol. Med.* **21**, 461–462 (2015).
83. Derks, S. & Diosdado, B. Personalized cancer medicine: next steps in the genomic era. *Cell. Oncol. (Dordr).* **38**, 1–2 (2015).
84. Veltman, J. A. & Lupski, J. R. From genes to genomes in the clinic. *Genome Med.* **7**, 78 (2015).
85. Gilbert, W. Towards a paradigm shift in biology. *Nature* **349**, 99 (1991).
86. Kumar, A. *et al.* PfalDB: An Integrated Drug Target and Chemical Database for Plasmodium falciparum. at <http://www.ingentaconnect.com/content/ben/cdt/2014/00000015/00000012/art00001>
>
87. Wassermann, A. M., Lounkine, E., Davies, J. W., Glick, M. & Camargo, L. M. The opportunities of mining historical and collective data in drug discovery. *Drug Discov. Today* **20**, 422–34 (2015).
88. Blundell, T. L. *et al.* Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 413–423 (2006).

89. de Beer, T. A. *et al.* Antimalarial drug discovery: in silico structural biology and rational drug design. *Infect. Disord. Drug Targets* **9**, 304–318 (2009).
90. Birkholtz, L.-M. *et al.* Integration and mining of malaria molecular, functional and pharmacological data: how far are we from a chemogenomic knowledge space? *Malar. J.* **5**, 110 (2006).
91. Setoain, J. *et al.* NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. *Nucleic Acids Res.* **43**, W193–9 (2015).
92. Alam, A. *et al.* Novel antimalarial drug targets: hope for new antimalarial drugs. *Expert Rev. Clin. Pharmacol.* **2**, 469–89 (2009).
93. Mehlotra, R. K., Henry-Halldin, C. N. & Zimmerman, P. A. Application of pharmacogenomics to malaria: a holistic approach for successful chemotherapy. *Pharmacogenomics* **10**, 435–449 (2009).
94. Allen, R. J. W. & Kirk, K. Cell volume control in the Plasmodium-infected erythrocyte. *Trends Parasitol.* **20**, 7–10 (2004).
95. Krugliak, M., Zhang, J. & Ginsburg, H. Intraerythrocytic Plasmodium falciparum utilizes only a fraction of the amino acids derived from the digestion of host cell cytosol for the biosynthesis of its proteins. *Mol. Biochem. Parasitol.* **119**, 249–256 (2002).
96. Qidwai, T. Hemoglobin degradation pathway of Plasmodium falciparum as antimalarial drug target. *Curr. Drug Targets* (2015).
97. Rosenthal, P. J., McKerrow, J. H., Aikawa, M., Nagasawa, H. & Leech, J. H. A malarial cysteine proteinase is necessary for hemoglobin degradation by Plasmodium falciparum. *J. Clin. Invest.* **82**, 1560–1566 (1988).
98. Goldberg, D. E. Plasmodial hemoglobin degradation: an ordered pathway in a specialized organelle. *Infect. Agents Dis.* **1**, 207–211 (1992).
99. Desai, P. V *et al.* Identification of novel parasitic cysteine protease inhibitors by use of virtual screening. 2. The available chemical directory. *J. Med. Chem.* **49**, 1576–84 (2006).
100. Desai, S. A., Bezrukov, S. M. & Zimmerberg, J. A voltage-dependent channel involved in nutrient uptake by red blood cells infected with the malaria parasite. *Nature* **406**, 1001–5 (2000).
101. Joet, T., Eckstein-Ludwig, U., Morin, C. & Krishna, S. Validation of the hexose transporter of Plasmodium falciparum as a novel drug target. *Proc. Natl. Acad. Sci.* **100**, 7476–7479 (2003).
102. Mok, S. *et al.* Artemisinin resistance in Plasmodium falciparum is associated with an altered temporal pattern of transcription. *BMC Genomics* **12**, 391 (2011).
103. Flohé, L., Hecht, H. J. & Steinert, P. Glutathione and trypanothione in parasitic hydroperoxide metabolism. *Free Radic. Biol. Med.* **27**, 966–84 (1999).
104. Rahlfs, S., Nickel, C., Deponte, M., Schirmer, R. H. & Becker, K. Plasmodium falciparum thioredoxins and glutaredoxins as central players in redox metabolism. *Redox Rep.* **8**, 246–50 (2003).

105. Bruns, C. M., Hubatsch, I., Ridderström, M., Mannervik, B. & Tainer, J. A. Human glutathione transferase A4-4 crystal structures and mutagenesis reveal the basis of high catalytic efficiency with toxic lipid peroxidation products. *J. Mol. Biol.* **288**, 427–39 (1999).
106. Gilberger, T.-W., Schirmer, R. H., Walter, R. D. & Müller, S. Deletion of the parasite-specific insertions and mutation of the catalytic triad in glutathione reductase from chloroquine-sensitive *Plasmodium falciparum* 3D7. *Mol. Biochem. Parasitol.* **107**, 169–179 (2000).
107. Meierjohann, S., Walter, R. D. & Müller, S. Glutathione synthetase from *Plasmodium falciparum*. *Biochem. J.* **363**, 833–8 (2002).
108. Sherman, I. W. Biochemistry of *Plasmodium* (malarial parasites). *Microbiol. Rev.* **43**, 453–95 (1979).
109. Gero, A. M. & O’Sullivan, W. J. Purines and pyrimidines in malarial parasites. *Blood Cells* **16**, 467–84; discussion 485–98 (1990).
110. Hassan, H. F. & Coombs, G. H. Purine and pyrimidine metabolism in parasitic protozoa. *FEMS Microbiol. Rev.* **4**, 47–83 (1988).
111. Dawson, P. A., Cochran, D. A., Emmerson, B. T. & Gordon, R. B. Inhibition of *Plasmodium falciparum* hypoxanthine-guanine phosphoribosyltransferase mRNA by antisense oligodeoxynucleotide sequence. *Mol. Biochem. Parasitol.* **60**, 153–6 (1993).
112. Krungkrai, J., Cerami, A. & Henderson, G. B. Pyrimidine biosynthesis in parasitic protozoa: purification of a monofunctional dihydroorotase from *Plasmodium berghei* and *Crithidia fasciculata*. *Biochemistry* **29**, 6270–5 (1990).
113. Rathod, P. K. & Reyes, P. Orotidylate-metabolizing enzymes of the human malarial parasite, *Plasmodium falciparum*, differ from host cell enzymes. *J. Biol. Chem.* **258**, 2852–5 (1983).
114. Olliaro, P. L. & Yuthavong, Y. An overview of chemotherapeutic targets for antimalarial drug discovery. *Pharmacol. Ther.* **81**, 91–110 (1999).
115. Le Roch, K. G. *et al.* A systematic approach to understand the mechanism of action of the bithiazolium compound T4 on the human malaria parasite, *Plasmodium falciparum*. *BMC Genomics* **9**, 513 (2008).
116. Waters, N. C. & Geyer, J. A. Cyclin-dependent protein kinases as therapeutic drug targets for antimalarial drug development. *Expert Opin. Ther. Targets* **7**, 7–17 (2003).
117. Klemba, M. & Goldberg, D. E. Biological roles of proteases in parasitic protozoa. *Annu. Rev. Biochem.* **71**, 275–305 (2002).
118. Li, X. *et al.* *Plasmodium falciparum* signal peptide peptidase is a promising drug target against blood stage malaria. *Biochem. Biophys. Res. Commun.* **380**, 454–9 (2009).
119. Rosenthal, P. J. Falcipains and other cysteine proteases of malaria parasites. *Adv. Exp. Med. Biol.* **712**, 30–48 (2011).
120. Debrabant, A. & Delplace, P. Leupeptin alters the proteolytic processing of P126, the major parasitophorous vacuole antigen of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **33**, 151–158 (1989).
121. Bailly, E., Jambou, R., Savel, J. & Jaureguiberry, G. *Plasmodium falciparum*:

- differential sensitivity in vitro to E-64 (cysteine protease inhibitor) and Pepstatin A (aspartyl protease inhibitor). *J. Protozool.* **39**, 593–599 (1992).
122. Adams, C. P. & Brantner, V. V. Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff. (Millwood)*. **25**, 420–8 (2006).
 123. Barnes, P. J. *et al.* Barriers to new drug development in respiratory disease. *Eur. Respir. J.* **45**, 1197–207 (2015).
 124. Pfister, D. G. The just price of cancer drugs and the growing cost of cancer care: oncologists need to be part of the solution. *J. Clin. Oncol.* **31**, 3487–9 (2013).
 125. Forbes, How Much Does Pharmaceutical Innovation Cost? (2011) A Look At 100 Companies. (2013) Available at <http://www.forbes.com/sites/matthewherper/2013/08/11/the-cost-of-inventing-a-new-drug-98-companies-ranked/>. (Accessed: 6th August 2015)
 126. Forbes, Pharma Be Aware: The Next Killers. (2011) Available at <http://www.forbes.com/sites/sciencebiz/2011/03/14/pharma-be-aware-the-next-killers/>. (Accessed: 16th August 2015)
 127. Forbes, The Cost Of Creating A New Drug Now \$5 Billion, Pushing Big Pharma To Change. (2013) Available at <http://www.forbes.com/sites/matthewherper/2013/08/11/how-the-staggering-cost-of-inventing-new-drugs-is-shaping-the-future-of-medicine/2/>. (Accessed: 12th August 2015)
 128. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–95 (2014).
 129. Kortagere, S. & Ekins, S. Troubleshooting computational methods in drug discovery. *J. Pharmacol. Toxicol. Methods* **61**, 67–75 (2014).
 130. Aguiar, A. C., Rocha, E. M., Souza, N. B., Franca, T. C. & Krettli, A. U. New approaches in antimalarial drug discovery and development: a review. *Mem. Inst. Oswaldo Cruz* **107**, 831–845 (2012).
 131. Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **152**, 38–52 (2007).
 132. Roses, A. D. Pharmacogenetics in drug discovery and development: a translational perspective. *Nat. Rev. Drug Discov.* **7**, 807–17 (2008).
 133. Bio Ventures for global health. Malaria Pipelines. (2014) Available at <http://www.bvgh.org/Current-Programs/Neglected-Disease-Product-Pipelines/Malaria-Pipelines.aspx>. (Accessed: 24th October 2015)
 134. Nwaka, S. & Ridley, R. G. Virtual drug discovery and development for neglected diseases through public-private partnerships. *Nat. Rev. Drug Discov.* **2**, 919–28 (2003).
 135. ChEMBL, Neglected Tropical Disease. Available at <https://www.ebi.ac.uk/chemblntd>. Accessed: 30th November 2015)
 136. Spangenberg, T. *et al.* The open access malaria box: a drug discovery catalyst for neglected diseases. *PLoS One* **8**, e62906 (2013).
 137. Schuldt, N. J. & Amalfitano, A. Malaria vaccines: focus on adenovirus based vectors. *Vaccine* **30**, 5191–5198 (2012).

138. Bhattarai, A. *et al.* Impact of artemisinin-based combination therapy and insecticide-treated nets on malaria burden in Zanzibar. *PLoS Med.* **4**, e309 (2007).
139. Griffin, J. T. *et al.* Reducing *Plasmodium falciparum* malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS Med.* **7**, e1000324 (2010).
140. Dondorp, A. M. *et al.* Artemisinin resistance: current status and scenarios for containment. *Nat. Rev.* **8**, 272–280 (2010).
141. Saralamba, S. *et al.* Intrahost modeling of artemisinin resistance in *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 397–402 (2011).
142. Liu, N. Insecticide Resistance in Mosquitoes: Impact, Mechanisms, and Research Directions. *Annu. Rev. Entomol.* **60**, 537–559 (2015).
143. Schäberle, T. F. & Hack, I. M. Overcoming the current deadlock in antibiotic research. *Trends Microbiol.* **22**, 165–7 (2014).
144. Wu, Y., Wang, X., Liu, X. & Wang, Y. Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. *Genome Res.* **13**, 601–616 (2003).
145. Gardiner, D. L. *et al.* *Plasmodium falciparum*: new molecular targets with potential for antimalarial drug development. *Expert Rev. Anti. Infect. Ther.* **7**, 1087–98 (2009).
146. Thovarai, V. in silico drug design of potential novel anti malarial agents. (2009).
147. McInnes, C. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* **11**, 494–502 (2007).
148. Sajid, M. & McKerrow, J. H. Cysteine proteases of parasitic organisms. *Mol. Biochem. Parasitol.* **120**, 1–21 (2002).
149. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38**, D227–33 (2010).
150. McKerrow, J. H., Caffrey, C., Kelly, B., Loke, P. & Sajid, M. Proteases in parasitic diseases. *Annu. Rev. Pathol.* **1**, 497–536 (2006).
151. McKerrow, J. H., Engel, J. C. & Caffrey, C. R. Cysteine protease inhibitors as chemotherapy for parasitic infections. *Bioorg. Med. Chem.* **7**, 639–644 (1999).
152. McKerrow, J. H. Development of cysteine protease inhibitors as chemotherapy for parasitic diseases: insights on safety, target validation, and mechanism of action. *Int. J. Parasitol.* **29**, 833–837 (1999).
153. Drew, M. E. *et al.* *Plasmodium* food vacuole plasmepsins are activated by falcipains. *J. Biol. Chem.* **283**, 12870–12876 (2008).
154. Chapman, H. a, Riese, R. J. & Shi, G. P. Emerging roles for cysteine proteases in human biology. *Annu. Rev. Physiol.* **59**, 63–88 (1997).
155. Pandey, K. C. & Dixit, R. Structure-function of falcipains: malarial cysteine proteases. *J. Trop. Med.* **2012**, 345195 (2012).
156. Erez, E., Fass, D. & Bibi, E. How intramembrane proteases bury hydrolytic reactions in the membrane. *Nature* **459**, 371–8 (2009).
157. Rosenthal, P. J. Cysteine proteases of malaria parasites. *Int. J. Parasitol.* **34**, 1489–

- 1499 (2004).
158. Stoka, V., Turk, B. & Turk, V. Lysosomal cysteine proteases: structural features and their role in apoptosis. *IUBMB Life* **57**, 347–353 (2005).
 159. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **40**, D343–50 (2012).
 160. Turk, V., Turk, B. & Turk, D. Lysosomal cysteine proteases: facts and opportunities. *EMBO J.* **20**, 4629–33 (2001).
 161. Hansen, G. *et al.* Structural basis for the regulation of cysteine-protease activity by a new class of protease inhibitors in Plasmodium. *Structure* **19**, 919–29 (2011).
 162. Ettari, R. *et al.* Falcipain-2 inhibitors. *Med. Res. Rev.* **30**, 136–167 (2010).
 163. Kerr, I. D. *et al.* Structures of falcipain-2 and falcipain-3 bound to small molecule inhibitors: implications for substrate specificity. *J. Med. Chem.* **52**, 852–857 (2009).
 164. Sijwali, P. S., Koo, J., Singh, N. & Rosenthal, P. J. Gene disruptions demonstrate independent roles for the four falcipain cysteine proteases of Plasmodium falciparum. *Mol. Biochem. Parasitol.* **150**, 96–106 (2006).
 165. Miller, L. H., Ackerman, H. C., Su, X. Z. & Wellems, T. E. Malaria biology and disease pathogenesis: insights for new treatments. *Nat. Med.* **19**, 156–167 (2013).
 166. Francis, S. E., Banerjee, R. & Goldberg, D. E. Biosynthesis and maturation of the malaria aspartic hemoglobinas plasmepsins I and II. *J. Biol. Chem.* **272**, 14961–14968 (1997).
 167. McKerrow, J. H., Sun, E., Rosenthal, P. J. & Bouvier, J. The proteases and pathogenicity of parasitic protozoa. *Annu. Rev. Microbiol.* **47**, 821–853 (1993).
 168. Rosenthal, P. J. Plasmodium falciparum: effects of proteinase inhibitors on globin hydrolysis by cultured malaria parasites. *Exp. Parasitol.* **80**, 272–281 (1995).
 169. Gamboa de Dominguez, N. D. & Rosenthal, P. J. Cysteine proteinase inhibitors block early steps in hemoglobin degradation by cultured malaria parasites. *Blood* **87**, 4448–4454 (1996).
 170. Pina-Vazquez, C., Reyes-Lopez, M., Ortiz-Estrada, G., de la Garza, M. & Serrano-Luna, J. Host-parasite interaction: parasite-derived and -induced proteases that degrade human extracellular matrix. *J. Parasitol. Res.* **2012**, 748206 (2012).
 171. Tardieux, I. & Menard, R. Migration of Apicomplexa across biological barriers: the Toxoplasma and Plasmodium rides. *Traffic* **9**, 627–635 (2008).
 172. Silvie, O., Franetich, J. F., Renia, L. & Mazier, D. Malaria sporozoite: migrating for a living. *Trends Mol. Med.* **10**, 91–97 (2004).
 173. Ishino, T., Yano, K., Chinzei, Y. & Yuda, M. Cell-passage activity is required for the malarial parasite to cross the liver sinusoidal cell layer. *PLoS Biol.* **2**, E4 (2004).
 174. Pandey, K. C., Singh, N., Arastu-Kapur, S., Bogyo, M. & Rosenthal, P. J. Falstatin, a cysteine protease inhibitor of Plasmodium falciparum, facilitates erythrocyte invasion. *PLoS Pathog.* **2**, e117 (2006).
 175. Rennenberg, A. *et al.* Exoerythrocytic Plasmodium parasites secrete a cysteine protease inhibitor involved in sporozoite invasion and capable of blocking cell death of

- host hepatocytes. *PLoS Pathog.* **6**, e1000825 (2010).
176. Hadley, T., Aikawa, M. & Miller, L. H. Plasmodium knowlesi: studies on invasion of rhesus erythrocytes by merozoites in the presence of protease inhibitors. *Exp. Parasitol.* **55**, 306–311 (1983).
 177. Wickham, M. E., Culvenor, J. G. & Cowman, A. F. Selective inhibition of a two-step egress of malaria parasites from the host erythrocyte. *J. Biol. Chem.* **278**, 37658–37663 (2003).
 178. Doyle, P. S. *et al.* The Trypanosoma cruzi protease cruzain mediates immune evasion. *PLoS Pathog.* **7**, e1002139 (2011).
 179. Rosenthal, P. J. & Nelson, R. G. Isolation and characterization of a cysteine proteinase gene of Plasmodium falciparum. *Mol. Biochem. Parasitol.* **51**, 143–152 (1992).
 180. Shenai, B. R., Sijwali, P. S., Singh, A. & Rosenthal, P. J. Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of Plasmodium falciparum. *J. Biol. Chem.* **275**, 29000–29010 (2000).
 181. Sijwali, P. S., Shenai, B. R., Gut, J., Singh, A. & Rosenthal, P. J. Expression and characterization of the Plasmodium falciparum haemoglobinase falcipain-3. *Biochem. J.* **360**, 481–489 (2001).
 182. Marco, M. & Coteron, J. M. Falcipain inhibition as a promising antimalarial target. *Curr. Top. Med. Chem.* **12**, 408–444 (2012).
 183. Kumar, A. *et al.* Falcipain-1, a Plasmodium falciparum cysteine protease with vaccine potential. *Infect. Immun.* **75**, 2026–34 (2007).
 184. Mane, U. R. *et al.* Falcipain inhibitors as potential therapeutics for resistant strains of malaria: a patent review. *Expert Opin. Ther. Pat.* **23**, 165–87 (2013).
 185. Rosenthal, P. J., Ring, C. S., Chen, X. & Cohen, F. E. Characterization of a Plasmodium vivax cysteine proteinase gene identifies uniquely conserved amino acids that may mediate the substrate specificity of malarial hemoglobinases. *J. Mol. Biol.* **241**, 312–316 (1994).
 186. Na, B. K. *et al.* Identification and biochemical characterization of vivapains, cysteine proteases of the malaria parasite Plasmodium vivax. *Biochem. J.* **378**, 529–538 (2004).
 187. Narayan, A. *et al.* Regulation of gene expression in Plasmodium falciparum. **102**, (2012).
 188. Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O. & Llinás, M. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.* **6**, e1001165 (2010).
 189. Bunnik, E. M. & Le Roch, K. G. PfAlba1: master regulator of translation in the malaria parasite. *Genome Biol.* **16**, 221 (2015).
 190. Sunil, S., Chauhan, V. S. & Malhotra, P. Distinct and stage specific nuclear factors regulate the expression of falcipains, Plasmodium falciparum cysteine proteases. *BMC Mol. Biol.* **9**, 47 (2008).
 191. Greenbaum, D. C. *et al.* A role for the protease falcipain 1 in host cell invasion by the human malaria parasite. *Science* **298**, 2002–2006 (2002).

192. Dahl, E. L. & Rosenthal, P. J. Biosynthesis, localization, and processing of falcipain cysteine proteases of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **139**, 205–212 (2005).
193. Goh, S. L., Goh, L. L. & Sim, T. S. Cysteine protease falcipain 1 in *Plasmodium falciparum* is biochemically distinct from its isozymes. *Parasitol. Res.* **97**, 295–301 (2005).
194. Salas, F., Fichmann, J., Lee, G. K., Scott, M. D. & Rosenthal, P. J. Functional expression of falcipain, a *Plasmodium falciparum* cysteine proteinase, supports its role as a malarial hemoglobinase. *Infect. Immun.* **63**, 2120–2125 (1995).
195. Wang, S. X. *et al.* Structural basis for unique mechanisms of folding and hemoglobin binding by a malarial protease. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11503–11508 (2006).
196. Pandey, K. C., Barkan, D. T., Sali, A. & Rosenthal, P. J. Regulatory elements within the prodomain of Falcipain-2, a cysteine protease of the malaria parasite *Plasmodium falciparum*. *PLoS One* **4**, e5694 (2009).
197. Sijwali, P. S., Brinen, L. S. & Rosenthal, P. J. Systematic optimization of expression and refolding of the *Plasmodium falciparum* cysteine protease falcipain-2. *Protein Expr. Purif.* **22**, 128–134 (2001).
198. Bromme, D., Nallaseth, F. S. & Turk, B. Production and activation of recombinant papain-like cysteine proteases. *Methods* **32**, 199–206 (2004).
199. Sijwali, P. S. & Rosenthal, P. J. Gene disruption confirms a critical role for the cysteine protease falcipain-2 in hemoglobin hydrolysis by *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4384–4389 (2004).
200. Pandey, K. C., Sijwali, P. S., Singh, A., Na, B. K. & Rosenthal, P. J. Independent intramolecular mediators of folding, activity, and inhibition for the *Plasmodium falciparum* cysteine protease falcipain-2. *J. Biol. Chem.* **279**, 3484–3491 (2004).
201. Hogg, T. *et al.* Structural and functional characterization of Falcipain-2, a hemoglobinase from the malarial parasite *Plasmodium falciparum*. *J. Biol. Chem.* **281**, 25425–25437 (2006).
202. Pandey, K. C. *et al.* The *Plasmodium falciparum* cysteine protease falcipain-2 captures its substrate, hemoglobin, via a unique motif. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 9138–9143 (2005).
203. Singh, A. & Rosenthal, P. J. Comparison of efficacies of cysteine protease inhibitors against five strains of *Plasmodium falciparum*. *Antimicrob. Agents Chemother.* **45**, 949–951 (2001).
204. Teixeira, C., Gomes, J. R. B. & Gomes, P. Falcipains, *Plasmodium falciparum* cysteine proteases as key drug targets against malaria. *Curr. Med. Chem.* **18**, 1555–1572 (2011).
205. Lee, B. J. *et al.* Antimalarial activities of novel synthetic cysteine protease inhibitors. *Antimicrob. Agents Chemother.* **47**, 3810–4 (2003).
206. Schulz, F. *et al.* Screening of protease inhibitors as antiplasmodial agents. Part I: Aziridines and epoxides. *ChemMedChem* **2**, 1214–24 (2007).

207. Rosenthal, P. J., Lee, G. K. & Smith, R. E. Inhibition of a *Plasmodium vinckei* cysteine proteinase cures murine malaria. *J. Clin. Invest.* **91**, 1052–6 (1993).
208. Wang, S. X. *et al.* The structure of chagasin in complex with a cysteine protease clarifies the binding mode and evolution of an inhibitor family. *Structure* **15**, 535–43 (2007).
209. Wang, J. *et al.* Structural features of falcipain-3 inhibitors: an in silico study. *Mol. Biosyst.* 2296–2310 (2013). doi:10.1039/c3mb70105k
210. Coteron, J. M. *et al.* Falcipain inhibitors: optimization studies of the 2-pyrimidinecarbonitrile lead series. *J. Med. Chem.* **53**, 6129–6152 (2010).
211. Aurrecochea, C. *et al.* PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37**, D539–43 (2009).
212. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **37**, D26–31 (2009).
213. Moreno-Hagelsieb, G. & Latimer, K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* **24**, 319–324 (2008).
214. Bernstein, F. C. *et al.* The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–42 (1977).
215. Kanzi, A. M. Falcipains As Malarial Drug Targets. (Rhodes University, 2013). doi:http://hdl.handle.net/10962/d1003842
216. Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics* **Chapter 5**, Unit 5.6 (2006).
217. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
218. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).
219. Lunter, G. *et al.* Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.* **18**, 298–309 (2008).
220. Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* **20**, 426–427 (2004).
221. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
222. Crooks, G., Hon, G., Chandonia, J. & Brenner, S. NCBI GenBank FTP Site\nWebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190 (2004).
223. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
224. Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Tastan Bishop, Ö. Analysis of Non-Peptidic Compounds as Potential Malarial Inhibitors against Plasmodial Cysteine Proteases via Integrated Virtual Screening Workflow. *J. Biomol. Struct. Dyn.* 1–72

- (2015).
225. Berg, J. M., Tymoczko, J. L. & Stryer, L. The Amino Acid Sequence of a Protein Determines Its Three-Dimensional Structure. (2002).
 226. Gillmor, S. a, Craik, C. S. & Fletterick, R. J. Structural determinants of specificity in the cysteine protease cruzain. *Protein Sci.* **6**, 1603–1611 (1997).
 227. Zhao, B. *et al.* Crystal structure of human osteoclast cathepsin K complex with E-64. *Nat. Struct. Biol.* **4**, 109–111 (1997).
 228. Sabnis, Y. A., Desai, P. V, Rosenthal, P. J. & Avery, M. A. Probing the structure of falcipain-3, a cysteine protease from *Plasmodium falciparum*: comparative protein modeling and docking studies. *Protein Sci.* **12**, 501–9 (2003).
 229. Desai, P. V & Avery, M. A. Structural characterization of vivapain-2 and vivapain-3, cysteine proteases from *Plasmodium vivax*: comparative protein modeling and docking studies. *J. Biomol. Struct. Dyn.* **21**, 781–90 (2004).
 230. Rzychon, M., Chmiel, D. & Stec-Niemczyk, J. Modes of inhibition of cysteine proteases. *Acta Biochim. Pol.* **51**, 861–873 (2004).
 231. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–9 (2001).
 232. Tachibana, S. *et al.* *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat. Genet.* **44**, 1051–1055 (2012).
 233. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–73 (2006).
 234. Anamika, Srinivasan, N. & Krupa, A. A genomic perspective of protein kinases in *Plasmodium falciparum*. *Proteins* **58**, 180–9 (2005).
 235. Wright, M. H. *et al.* Validation of N-myristoyltransferase as an antimalarial drug target using an integrated chemical biology approach. *Nat. Chem.* **6**, 112–21 (2014).
 236. Goldston, A. M., Sharma, A. I., Paul, K. S. & Engman, D. M. Acylation in trypanosomatids: an essential process and potential drug target. *Trends Parasitol.* **30**, 350–60 (2014).
 237. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–32 (1982).
 238. Dror, R. O., Jensen, M. Ø., Borhani, D. W. & Shaw, D. E. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *J. Gen. Physiol.* **135**, 555–62 (2010).
 239. Mobley, D. L. & Klimovich, P. V. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.* **137**, 230901 (2012).
 240. Martín-García, F., Papaleo, E., Gomez-Puertas, P., Boomsma, W. & Lindorff-Larsen, K. Comparing molecular dynamics force fields in the essential subspace. *PLoS One* **10**, e0121114 (2015).
 241. Bond, P. J., Holyoake, J., Ivetac, A., Khalid, S. & Sansom, M. S. P. Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J. Struct. Biol.*

- 157, 593–605 (2007).
242. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 826–843 (2011).
 243. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12562–6 (2002).
 244. Schlacken, H. Molecular-dynamics simulation of statistical-mechanical systems. Proceedings of the Enrico Fermi International Summer School of Physics, 23 July–2 August, 1985. Hg. von G. CICCOTTI und W. G. HOOVER. Amsterdam: North-Holland Elsevier Science Publisher 1987. *Acta Polym.* **39**, 151–152 (1988).
 245. Alder, B. J. & Wainwright, T. E. Studies in molecular dynamics. General method. *J. Chem. Phys.* **31**, 459 (1959).
 246. Levitt, M. & Warshel, A. Computer simulation of protein folding. *Nature* **253**, 694–8 (1975).
 247. Cooper, A. Thermodynamic fluctuations in protein molecules. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 2740–1 (1976).
 248. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).
 249. Case, D. A. & Karplus, M. Dynamics of ligand binding to heme proteins. *J. Mol. Biol.* **132**, 343–68 (1979).
 250. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–88 (1982).
 251. Northrup, S. H., Pear, M. R., Lee, C. Y., McCammon, J. A. & Karplus, M. Dynamical theory of activated processes in globular proteins. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 4035–9 (1982).
 252. Brooks, B. & Karplus, M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 6571–5 (1983).
 253. Levitt, M. & Sharon, R. Accurate simulation of protein dynamics in solution. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7557–61 (1988).
 254. Guilbert, C., Perahia, D. & Mouawad, L. A method to explore transition paths in macromolecules. Applications to hemoglobin and phosphoglycerate kinase. *Comput. Phys. Commun.* **91**, 263–273 (1995).
 255. Israelachvili, J. & Wennerström, H. Role of hydration and water structure in biological and colloidal interactions. *Nature* **379**, 219–25 (1996).
 256. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–25 (1997).
 257. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–94 (1998).
 258. Lu, H., Israilewitz, B., Krammer, A., Vogel, V. & Schulten, K. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys. J.* **75**,

- 662–71 (1998).
259. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
 260. Bernèche, S. & Roux, B. Energetics of ion conduction through the K⁺ channel. *Nature* **414**, 73–7 (2001).
 261. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–5 (2008).
 262. Shaw, D. E. *et al.* Millisecond-scale molecular dynamics simulations on Anton. in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09 1* (ACM Press, 2009).
 263. Galindo-Murillo, R., Roe, D. R. & Cheatham, T. E. Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta* **1850**, 1041–58 (2015).
 264. Bou-Rabee, N. Time integrators for molecular dynamics. *Entropy* **16**, 138–162 (2014).
 265. Jung, J. *et al.* GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 310–323 (2015).
 266. Witze, A. Joint effort nabs next wave of US supercomputers. *Nature* **515**, 324 (2014).
 267. Gara, a. *et al.* Overview of the Blue Gene/L system architecture. *IBM J. Res. Dev.* **49**, 195–212 (2005).
 268. Shaw, D. E. *et al.* Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. in *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis* 41–53 (IEEE, 2014). doi:10.1109/SC.2014.9
 269. Ma, B., Shatsky, M., Wolfson, H. J. & Nussinov, R. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **11**, 184–97 (2002).
 270. Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2**, 527–541 (2003).
 271. Kumar, S., Ma, B., Tsai, C. J., Wolfson, H. & Nussinov, R. Folding funnels and conformational transitions via hinge-bending motions. *Cell Biochem. Biophys.* **31**, 141–64 (1999).
 272. Raval, A., Piana, S., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* **80**, 2071–9 (2012).
 273. Schames, J. R. *et al.* Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* **47**, 1879–81 (2004).
 274. Grant, B. J. *et al.* Novel allosteric sites on Ras for lead generation. *PLoS One* **6**, e25711 (2011).
 275. Durrant, J. D., Keränen, H., Wilson, B. A. & McCammon, J. A. Computational identification of uncharacterized cruzain binding sites. *PLoS Negl. Trop. Dis.* **4**, e676

- (2010).
276. Satoh, M. *et al.* Multiple binding modes of a small molecule to human Keap1 revealed by X-ray crystallography and molecular dynamics simulation. *FEBS Open Bio* **5**, 557–70 (2015).
 277. Chen, L., Zheng, Q.-C. & Zhang, H.-X. Insights into the effects of mutations on Cren7-DNA binding using molecular dynamics simulations and free energy calculations. *Phys. Chem. Chem. Phys.* **17**, 5704–11 (2015).
 278. Ehrenfest, P. Bemerkung über die angenäherte Gültigkeit der klassischen Mechanik innerhalb der Quantenmechanik. *Zeitschrift für Phys.* **45**, 455–457 (1927).
 279. Karplus, M. & Petsko, G. a. Molecular dynamics simulations in biology. *Nature* **347**, 631–639 (1990).
 280. González, M. a. Force fields and molecular dynamics simulations. *École thématique la Société Française la Neutron.* **12**, 169–200 (2011).
 281. Berk Hess, David van der Spoel, E. L. Gromacs User Manual. *Dep. Biophys. Chem. ...* 312 (2014).
 282. Lifson, S. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *J. Chem. Phys.* **49**, 5116 (1968).
 283. Jorgensen, W. L. & Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 6665–70 (2005).
 284. Models, M. M., Philosophy, A. & Kollman, P. A. Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules. *Acc. Chem. Res.* **29**, 461–469 (1996).
 285. Weisstein, E. W. Taylor Series. at <<http://mathworld.wolfram.com/TaylorSeries.html>>
 286. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–52 (2002).
 287. Makarov, V., Pettitt, B. M. & Feig, M. Solvation and Hydration of Proteins and Nucleic Acids: A Theoretical View of Simulation and Experiment. *Acc. Chem. Res.* **35**, 376–384 (2002).
 288. Cheatham, T. E. & Kollman, P. A. Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* **51**, 435–71 (2000).
 289. Xia, B., Tsui, V., Case, D. A., Dyson, H. J. & Wright, P. E. Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water. *J. Biomol. NMR* **22**, 317–31 (2002).
 290. Orozco, M. & Luque, F. J. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev.* **100**, 4187–4226 (2000).
 291. Tsui, V. & Case, D. A. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* **56**, 275–91
 292. Simonson, T. Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol.* **11**, 243–252 (2001).

293. Mark, P. & Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **105**, 9954–9960 (2001).
294. Domínguez, J. N. *et al.* Synthesis and evaluation of new antimalarial phenylurenyl chalcone derivatives. *J. Med. Chem.* **48**, 3654–8 (2005).
295. Batra, S., Sabnis, Y. A., Rosenthal, P. J. & Avery, M. A. Structure-based approach to falcipain-2 inhibitors: synthesis and biological evaluation of 1,6,7-trisubstituted dihydroisoquinolines and isoquinolines. *Bioorg. Med. Chem.* **11**, 2293–2299 (2003).
296. Chipeleme, A., Gut, J., Rosenthal, P. J. & Chibale, K. Synthesis and biological evaluation of phenolic Mannich bases of benzaldehyde and (thio)semicarbazone derivatives against the cysteine protease falcipain-2 and a chloroquine resistant strain of *Plasmodium falciparum*. *Bioorg. Med. Chem.* **15**, 273–82 (2007).
297. Chiyanzu, I. *et al.* Synthesis and evaluation of isatins and thiosemicarbazone derivatives against cruzain, falcipain-2 and rhodesain. *Bioorg. Med. Chem. Lett.* **13**, 3527–3530 (2003).
298. Greenbaum, D. C. *et al.* Synthesis and Structure - Activity Relationships of Parasitocidal Thiosemicarbazone Cysteine Protease Inhibitors against *Plasmodium falciparum*, *Trypanosoma brucei*, and *Trypanosoma cruzi*. *J. Med. Chem.* **47**, 3212–3219 (2004).
299. Ngo, L. T., Okogun, J. I. & Folk, W. R. 21st century natural product research and drug development and traditional medicines. *Nat. Prod. Rep.* **30**, 584–92 (2013).
300. Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Bishop, Ö. T. Structure Based Docking and Molecular Dynamic Studies of Plasmodial Cysteine Proteases against a South African Natural Compound and its Analogs. *Nat. Publ. Gr.* 1–12 (2016).
301. Brown, D. K., Penkler, D. L., Musyoka, T. M. & Bishop, Ö. T. JMS: An Open Source Workflow Management System and Web-Based Cluster Front-End for High Performance Computing. *PLoS One* **10**, e0134273 (2015).
302. Wang, J., Wang, W., Kollman, P. a & Case, D. a. Antechamber, An Accessory Software Package For Molecular Mechanical Calculations. *J. Comput. Chem.* **25**, 1157–1174 (2005).
303. Mayne, C. G., Gumbart, J. C. & Tajkhorshid, E. The Force Field Toolkit: Software for the Parameterization of Small Molecules from First Principles. *Biophys. J.* **104**, 31a (2013).
304. van Aalten, D. M. *et al.* PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *J. Comput. Aided. Mol. Des.* **10**, 255–62 (1996).
305. Malde, A. K. *et al.* An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **7**, 4026–4037 (2011).
306. Sousa da Silva, A. W. & Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interface. *BMC Res. Notes* **5**, 367 (2012).
307. Mulliken, R. S. A New Electroaffinity Scale; Together with Data on Valence States and on Valence Ionization Potentials and Electron Affinities. *J. Chem. Phys.* **2**, 782 (1934).

308. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–54 (2013).
309. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
310. Parrinello, M. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182 (1981).
311. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
312. Petersen, H. G. Accuracy and efficiency of the particle mesh Ewald method. *J. Chem. Phys.* **103**, 3668 (1995).
313. Laskowski, R. A. & Swindells, M. B. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* **51**, 2778–2786 (2011).
314. Centre for high performance computing. at <<http://www.chpc.ac.za/>>
315. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. **46**, 3–26 (2001).
316. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–23 (2002).
317. Keller, T. H., Pichota, A. & Yin, Z. A practical view of ‘druggability’. *Curr. Opin. Chem. Biol.* **10**, 357–61 (2006).
318. DruLiTO, Drug Likeness Tool. Available at: http://www.niper.gov.in/pi_dev_tools/DruLiToWeb/DruLiTo_index.html. (Accessed: 24th November 2015)
319. O’Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).
320. Lobanov, M. I., Bogatyreva, N. S. & Galzitskaia, O. V. Radius of gyration is indicator of compactness of protein structure. *Mol. Biol. (Mosk)*. **42**, 701–706 (2008).
321. Chan, M. K., Mukund, S., Kletzin, A., Adams, M. W. & Rees, D. C. Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science* **267**, 1463–9 (1995).
322. Carlson, H. A. & McCammon, J. A. Accommodating Protein Flexibility in Computational Drug Design. *Mol. Pharmacol.* **57**, 213–218 (2000).
323. Davies-Coleman, M. T. & Beukes, D. R. Ten years of marine natural products research at Rhodes University. *S. Afr. J. Sci.* **100**, 539–544 (2004).
324. Ursu, O., Rayan, A., Goldblum, A. & Oprea, T. I. Understanding drug-likeness. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 760–781 (2011).
325. Lizana, L., Bauer, B. & Orwar, O. Controlling the rates of biochemical reactions and signaling networks by shape and volume changes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4099–104 (2008).
326. Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.* **5**, 821–34 (2006).

327. Frederick, K. K., Marlow, M. S., Valentine, K. G. & Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **448**, 325–9 (2007).
328. Copeland, R. A. *Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis*. (2000). at <<http://elib.tic.edu.vn:8080/dspace/handle/123456789/5467>>
329. Wang, J., Deng, Y. & Roux, B. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.* **91**, 2798–814 (2006).
330. Singh, N. & Warshel, A. Absolute binding free energy calculations: on the accuracy of computational scoring of protein-ligand interactions. *Proteins* **78**, 1705–23 (2010).
331. Deng, Y. & Roux, B. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B* **113**, 2234–46 (2009).
332. Izadi, S., Aguilar, B. & Onufriev, A. V. Protein–Ligand Electrostatic Binding Free Energies from Explicit and Implicit Solvation. *J. Chem. Theory Comput.* **11**, 4450–4459 (2015).
333. Muegge, I. & Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **42**, 791–804 (1999).
334. Jain, A. N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided. Mol. Des.* **10**, 427–440 (1996).
335. Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided. Mol. Des.* **8**, 243–256 (1994).
336. Feierberg, I., Feierberg, I., Luzhkov, V. B. & Luzhkov, V. B. Free Energy Calculations and Ligand Binding. *Adv. Protein Chem.* 123–158 (2003). doi:10.1016/S0065-3233(03)66004-3
337. Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **3**, 300–313 (1935).
338. Warshel, A. Dynamics of Enzymatic Reactions. *Proc. Natl. Acad. Sci.* **81**, 444–448 (1984).
339. Tembe, B. L. & McCammon, J. A. Ligand-Receptor Interactions. *Comput. Chem.* **8**, 281–283 (1984).
340. Kumari, R., Kumar, R. & Lynn, A. g_mmpbsa-A GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* **54**, 1951–1962 (2014).
341. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A. & Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate–DNA Helices. *J. Am. Chem. Soc.* **120**, 9401–9409 (1998).
342. Gershell, L. J. & Atkins, J. H. A brief history of novel drug discovery technologies. *Nat. Rev. Drug Discov.* **2**, 321–7 (2003).
343. Oprea, T. I., Zamora, I. & Ungell, A.-L. Pharmacokinetically based mapping device for chemical space navigation. *J. Comb. Chem.* **4**, 258–66
344. Fox, S. *et al.* High-throughput screening: update on practices and success. *J. Biomol.*

- Screen.* **11**, 864–9 (2006).
345. Mayr, L. M. & Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **9**, 580–8 (2009).
346. Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **1**, 882–94 (2002).
347. Geppert, H., Vogt, M. & Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **50**, 205–16 (2010).
348. Johnson, M., Basak, S. & Maggiora, G. A characterization of molecular similarity methods for property prediction. *Math. Comput. Model.* **11**, 630–634 (1988).
349. Hert, J., Irwin, J. J., Laggner, C., Keiser, M. J. & Shoichet, B. K. Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol.* **5**, 479–83 (2009).
350. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–49 (2004).
351. Ferreira, L., dos Santos, R., Oliva, G. & Andricopulo, A. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **20**, 13384–13421 (2015).
352. Yuriev, E. & Ramsland, P. A. Latest developments in molecular docking: 2010–2011 in review. *J. Mol. Recognit.* **26**, 215–39 (2013).
353. Plewczynski, D., Łażniewski, M., Augustyniak, R. & Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **32**, 742–55 (2011).
354. Penkler, D. L. In silico analysis of human Hsp90 for the identification of novel anti-cancer drug target sites and natural compound inhibitors. (Rhodes University, 2015). doi:<http://hdl.handle.net/10962/d1018938>
355. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J. & Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **153 Suppl**, S7–26 (2008).
356. Morris, G. M. *et al.* Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
357. Ewing, T. J. A., Makino, S., Skillman, A. G. & Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided. Mol. Des.* **15**, 411–428
358. Kramer, B., Rarey, M. & Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **37**, 228–41 (1999).
359. Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **46**, 499–511 (2003).
360. McGann, M. FRED and HYBRID docking performance on standardized datasets. *J. Comput. Aided. Mol. Des.* **26**, 897–906 (2012).
361. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D.

- Improved protein-ligand docking using GOLD. *Proteins* **52**, 609–23 (2003).
362. Hetényi, C. & van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **11**, 1729–37 (2002).
363. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–61 (2010).
364. Mohan, V., Gibbs, A. C., Cummings, M. D., Jaeger, E. P. & DesJarlais, R. L. Docking: successes and challenges. *Curr. Pharm. Des.* **11**, 323–333 (2005).
365. Totrov, M. & Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **18**, 178–84 (2008).
366. Taylor, R. D., Jewsbury, P. J. & Essex, J. W. FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *J. Comput. Chem.* **24**, 1637–56 (2003).
367. M.W. Schmidt J.A. Boatz, S.T. Elbert, M.S. Gordon, J.H. Jensen, S.Koseki, N.Matsunaga, K.A.Nguyen, S.Su, T.L.Windus, M.Dupuis, J.A.Montgomery, K. k. B. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **14**, 1347–1363 (1993).
368. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res.* **43**, W443–W447 (2015).
369. van Heerden, F. R. *et al.* An appetite suppressant from Hoodia species. *Phytochemistry* **68**, 2545–53 (2007).
370. Mimaki, Y. *et al.* Cholestane glycosides with potent cytostatic activities on various tumor cells from *Ornithogalum saundersiae* bulbs. *Bioorg. Med. Chem. Lett.* **7**, 633–636 (1997).
371. de Oliveira, M. *et al.* Antimalarial Activity of 4-Metoxychalcones: Docking Studies as Falcipain/Plasmeprin Inhibitors, ADMET and Lipophilic Efficiency Analysis to Identify a Putative Oral Lead Candidate. *Molecules* **18**, 15276–15287 (2013).
372. Ettari, R. *et al.* Synthesis and molecular modeling studies of derivatives of a highly potent peptidomimetic vinyl ester as falcipain-2 inhibitors. *ChemMedChem* **7**, 1594–600 (2012).
373. Ang, K. K. H. *et al.* Mining a cathepsin inhibitor library for new antiparasitic drug leads. *PLoS Negl. Trop. Dis.* **5**, e1023 (2011).
374. Teixeira, C., Gomes, J. R. B., Couesnon, T. & Gomes, P. Molecular docking and 3D-quantitative structure activity relationship analyses of peptidyl vinyl sulfones: Plasmodium Falciparum cysteine proteases inhibitors. *J. Comput. Aided. Mol. Des.* **25**, 763–775 (2011).