Structural bioinformatics studies and tool development related to drug discovery

A thesis submitted in fulfilment of the requirement for the degree

of

DOCTOR OF PHILOSOPHY
IN BIOINFORMATICS

of

RHODES UNIVERSITY, SOUTH AFRICA

Department of Biochemistry and Microbiology
Faculty of Science

by

Rowan Hatherley
June 2015

ABSTRACT

This thesis is divided into two distinct sections which can be combined under the broad umbrella of structural bioinformatics studies related to drug discovery. The first section involves the establishment of an online South African natural products database. Natural products (NPs) are chemical entities synthesised in nature and are unrivalled in their structural complexity, chemical diversity, and biological specificity, which has long made them crucial to the drug discovery process. South Africa is rich in both plant and marine biodiversity and a great deal of research has gone into isolating compounds from organisms found in this country. However, there is no official database containing this information, making it difficult to access for research purposes. This information was extracted manually from literature to create a database of South African natural products. In order to make the information accessible to the general research community, a website, named "SANCDB", was built to enable compounds to be quickly and easily searched for and downloaded in a number of different chemical formats. The content of the database was assessed and compared to other established natural product databases. Currently, SANCDB is the only database of natural products in Africa with an online interface.

The second section of the thesis was aimed at performing structural characterisation of proteins with the potential to be targeted for antimalarial drug therapy. This looked specifically at 1) The interactions between an exported heat shock protein (Hsp) from *Plasmodium falciparum* (*P. falciparum*), PfHsp70-x and various host and exported parasite J proteins, as well as 2) The interface between PfHsp90 and the heat shock organising protein (PfHop). The PfHsp70-x:J protein study provided additional insight into how these two proteins potentially interact. Analysis of the PfHsp90:PfHop also provided a structural insight into the interaction interface between these two proteins and identified residues that could be

targeted due to their contribution to the stability of the Hsp90:Hop binding complex and differences between parasite and human proteins. These studies inspired the development of a homology modelling tool, which can be used to assist researchers with homology modelling, while providing them with step-by-step control over the entire process.

This thesis presents the establishment of a South African NP database and the development of a homology modelling tool, inspired by protein structural studies. When combined, these two applications have the potential to contribute greatly towards *in silico* drug discovery research.

DECLARATION

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

_____

_____

This _____ day of _____ 2015

ACKNOWLEDGEMENTS

I would like to acknowledge the following people for their contributions to this work:

Firstly, and certainly the biggest thank you of all, to my supervisor, Prof. Özlem Taştan Bishop. I cannot say I made things easy for her, yet somehow she never gave up on me. For always believing in me and pushing me to believe in myself and my work. For her kindness, her care, her friendship and for providing me with more opportunities than I could ever image. For everything she has done for me, I truly cannot thank her enough. It means more to me than I could ever say.

Dr. Kevin Lobb, from the Rhodes Chemistry department, for his chemical insight into the SANCDB database and for his work assigning classifications to each compound and checking their 3D structures.

Prof. Gerard Kleywegt for hosting my Summer training at the European Bioinformatics Institute (EBI) and to Dr Sameer Velankar for supervising the work I did there. To these two and all the people I met and worked with at EBI, a most sincere thank you for such an amazing opportunity. I gained so much from this experience.

Prof. Greg Blatch and Dr. Eva-Rachele Pesce, for introducing me to the world of molecular chaperones, specifically PfHsp70-x, back when I worked in the wet lab. This was a time when I learned what passion for research was, largely thanks to these two individuals.

David Brown, for providing guidance with the development of the database, website and homology modelling tool, and for the most part helping to develop the 'informatics' side of my bioinformatics skill set.

The depositors working on the SANCDB project, Thommas Musyoka, David Penkler, Ngonidzashe Faya. Also to David Penkler for helping out with some of the proof-reading towards the end of the project, especially when English failed me.

Members of the RUBi lab, past and present. For their personal insight into my work and occasionally allowing me bounce ideas of your heads when I felt I needed someone to talk to.

Personal acknowledgements

The support I received for this project came in many forms. The following contributions were separate to those of academic or financial value, but were still incredibly important to me and I honestly would not have made it through this process without these people in my life.

To all my friends, family and colleagues for their support during my PhD.

The members of my "tea club" - Dr. Jacqueline van Marwijk, Sagaran Abboo, Margot Brooks, Kally Fitzgerald and Joan Miles. For providing me with their friendship in the early days of my PhD, when this was a rare commodity in my life.

My parents, Susan and Gavin, and Dee. For their continued support and for listening and nodding politely when I explain my work to them.

Throughout the years of my PhD, especially towards the end, I met some truly amazing people, whose friendship helped me to find strength when times were rough. Though I cannot express the ways in which each of these individuals impacted on my life (without spanning these sentiments over several more pages), I would like to make a special mention of them - Benjamin Miller, David Brown, David Penkler, Douglas Eastment, Jessica Vercueil, Kally Fitzgerald, Kéri Werth and Sharleigh Talbot - A huge thank you for helping me in more ways than you will ever know.

Above all else, to my sister Holly. My strength and eternal inspiration – you are forever in my heart.

Acknowledgement of funding

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

This work is partially supported by the National Institutes of Health Common Fund under grant number U41HG006941 to H3ABioNet.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

RESEARCH OUTPUTS

Conference Poster Presentations

Hatherley, RA, Blatch GL & Tastan Bishop Ö. **"*Plasmodium falciparum* Hsp70-x: An Atypical Cytosolic Heat Shock Protein."** *Joint Conference of the South African Genetics and South African Society for Bioinformatics and Computational Biology*, Stellenbosch, South Africa, 10-12 September 2012.

Conference Oral Presentations

Hatherley, RA, Brown, DK, Musyoka, T, Faya, N, Penkler, D, & Taştan Bishop, Ö. **"Design and Development of a South African Natural Compounds Database."** *Joint SASBi-SAGS Congress*, Kwalata Game Ranch, Tshwane, South Africa, 23-26 September 2014.

Hatherley, RA, Brown, DK, Musyoka, T, Faya, N, Penkler, D, & Taştan Bishop, Ö. **"Design and Development of a South African Natural Compounds Database."** *The 6th Interdisciplinary Post Graduate Conference*, Rhodes University, 8-10 October 2014.

Hatherley, RA, Brown, DK, Musyoka, T, Faya, N, Penkler, D, & Taştan Bishop, Ö. **"Design and Development of a South African Natural Compounds Database."** *Symposium on Chemico- and Biomedicinal Research*, Rhodes University, 27 October, 2014

Publications

Hatherley R, Blatch GL, Tastan Bishop Ö: ***Plasmodium falciparum* Hsp70-x: A heat shock protein at the host-parasite interface**. *J Biomol Struct Dyn* 2014, **32**:1766–1779.

Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, Dana JM, Fernandez Montecelo MA, Van Ginkel G, Gore SP, Haslam P, Hatherley R, Hendrickx PMS, Hirshberg M, Lagerstedt I, Mir S, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Rinaldi L, Sahni G, Sanz-García E, Sen S, Slowley RA, Velankar S, Wainwright ME, Kleywegt GJ: **PDBe: Protein data bank in Europe**. *Nucleic Acids Res* 2014, **42**(Database Issue):285–291.

Hatherley R, Clitheroe C, Faya N, Tastan Bishop Ö: ***Plasmodium falciparum* Hop: Detailed analysis on complex formation with Hsp70 and Hsp90**. *Biochem Biophys Res Commun* 2015, **456**:440–445.

Hatherley R, Brown, DK, Musyoka, T, Penkler, D, Faya, N, Lobb, KA, Taştan Bishop, Ö: **SANCDB: A South African Natural Compound Database.** *J Cheminform*, in press.

Contributions to Publications

Hatherley *et al*., 2014: ***Plasmodium falciparum*** **Hsp70-x: A heat shock protein at the host-parasite interface.**

For this article, I performed all experiments and wrote the first draft of the manuscript, under the guidance of Prof. GL Blatch and Prof. Ö Taştan Bishop.

Gutmanas *et al*., 2014: **PDBe: Protein data bank in Europe**.

This article was a part of a Summer internship at the European Bioinformatics Institute (EBI). My contribution included working on a project for the EBI, which involved automated identification of small molecules, present in PDB structures that may have biological significance and why.

Hatherley *et al*., 2015: ***Plasmodium falciparum*** **Hop: Detailed analysis on complex formation with Hsp70 and Hsp90.**

My contributions to this article involved repeating initial structural studies performed by a previous student, C Clitheroe, as well as performing subsequent structural analysis using my own scripts and wrote the structural sections of the manuscript.

Hatherley *et al*., (in press): **SANCDB: A South African Natural Compound Database.**

For this article, I created the database and website with the assistance of D Brown and under the guidance of Dr KA Lobb and Prof. Ö Taştan Bishop. I was involved in the compound uploading process, along with T Musyoka, D Penkler and N Faya, though curated work done by these individuals. I also wrote the first draft of the manuscript.

LIST OF ABBREVIATIONS

Abbreviation	Description
ACTs	Artemisinin-based combination therapies
ADMET	Absorption, distribution, metabolism, excretion and toxicity
Aha1	Activation of Hsp90 ATPase protein 1
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
CAS	Chemical Abstracts Service
CASP	Critical Assessment of protein Structure Prediction
CASRN	CAS Registry Number
ChEBI	Chemical Entities of Biological Interest
CNPD	Chinese Natural Product Database
CSD	Cambridge Structural Database
DMT	Department of Traditional Medicine
DOPE	Discrete optimized protein energy
DSG	15-deoxyspergualin
E. coli	*Escherichia coli*
ER	Endoplasmic reticulum
FFT	Fast Fourier transform
GSK	GlaxoSmithKline
Hip	Hsp70-interacting protein
HMDB	Human Metabolome Database
HMM	Hidden Markov model
Hop	Heat shock organizing protein
HPD	Histidine-Proline-Aspartic acid
Hsp110	110 kDa heat shock protein
Hsp70	70 kDa heat shock protein
Hsp90	90 kDa heat shock protein
HTML	Hyper Text Markup Language
HTS	High throughput screening
IPT	Intermittent preventive treatment
ITN	Insect-treated net
MDR	Multiple drug resistance
Molfile	Molecule file
MP	Misfolded peptide
MQAPs	Model quality assessment programs
MSAs	Multiple sequence alignments
MTV	Model-Template-View
NBD	Nucleotide binding domain
NCBI	National Center for Biotechnology Information
NCE	New chemical entity
NEF	Nucleotide exchange factor
NMR	Nuclear magnetic resonance

NOE	Nuclear Overhauser effect
NRB	Number of rotatable bonds
NRF	National Research Foundation
P. falciparum	*Plasmodium falciparum*
PDB	Protein Data Bank
PDF	Probability density function
PfHop	*Plasmodium falciparum* Hop
PfHSBP	*Plasmodium falciparum* heat shock factor binding protein
PfHsp70	*Plasmodium falciparum* 70 kDa heat shock protein
PfHsp90	*Plasmodium falciparum* 90 kDa heat shock protein
PfJ	*Plasmodium falciparum* J Protein
ProSA	Protein Structure Analysis
PSI-BLAST	Position-Specific Iterated BLAST
QSAR	Quantitative structure activity relationships
RBCs	Red blood cells
RMSD	Root mean square deviation
RSC	Royal Society of Chemistry
SANBI	South African National Biodiversity Institute
TCM	Traditional Chinese medicine
TM	Traditional medicine
UI	User interface
WHO	World Health Organization

THESIS STRUCTURE

This thesis has been divided into two independent sections, which are united under the common theme of drug discovery. The first part of the thesis, which consists of Chapters 1, 2 and 3, is centred around working with natural products data. More specifically, the establishment of a freely accessible, online database of compounds, isolated from organisms found in South Africa.

Chapter 1 gives a general introduction to natural products and why such a great deal of research attention has been given to them. This looks into their use in traditional medicines and how these have contributed to our modern day understanding of drugs and drug development, with specific examples of drugs that have been discovered as a direct result of their use, or the use of their source organisms, in traditional medicine. The vast biodiversity of South Africa is then described, along with the pharmaceutical potential of both plant and marine life from this part of the world. Finally, the reasoning for the research undertaken in this part of the thesis is given. Simply put, in spite of a great deal of research that has gone into isolating and characterising natural products from South Africa, there is currently no official database containing this information, nor a resource which allows researchers to easily find and access these compounds or work with their chemical data. This problem is addressed in Chapters 2 and 3, firstly by constructing and establishing a database of South African natural products, then by creating on online resource which makes this information easily and freely accessible.

Chapter 2 describes the design and layout of a South African natural products database, including the collection of compound data from literature, as well as other information considered relevant for further study of these compounds. To contextualise this, the introduction to this chapter reviews a number of well-established and prominent chemical

databases, to show both the diversity of focuses employed by these databases and the magnitude of data contained within. The process of building the database, collecting information and addressing data curation and quality control are described. The content of the database is discussed and compared to other natural products databases. Plans for further development of the database are also discussed.

Chapter 3 describes the construction of an online interface for the database which has been named "SANCDB". The background given in this chapter focuses more on the value of making such data publically accessible and the means by which this can be achieved. The development of the website is described along with the process by which users may search for information, download content or follow links to access additional data. Future developments to the site are also discussed, largely within the context of the expansion of the database itself.

Section 2 is focused around protein structural bioinformatics, with Chapter 4 giving a general introduction to this part of the thesis. An overview of the study of protein structures is given with reference to gaining a mechanistic insight to better understand protein function. The efforts of structural genomics are reviewed, as are the challenges of obtaining experimental structures for proteins and how *in silico* approaches can be used to study proteins with structures that have not yet been solved by experimental means. A brief overview of the work performed in this part of the thesis is given. Primarily, this includes structural studies involving malarial heat shock proteins (Hsps) and interactions with their cochaperones (Chapter 5), which led to a secondary aim of assisting with future structural studies by providing a means to quickly and reliably perform homology modelling of target proteins (Chapter 6).

Chapter 5 focuses on the structural characterisation of Hsps from *Plasmodium falciparum* (*P. falciparum*) and their human counterparts. Specifically, these include Hsp70s, Hsp90s, J proteins and the Hsp70/Hsp90 organising protein, Hop. At a broad perspective, Hsp70s are molecular chaperones involved in protein folding and preventing aggregation, but also have other functions. Their activity is both assisted and mediated by J proteins and in some cases, the final maturation of peptide substrates involves transfer to Hsp90, which is facilitated by the cochaperone, Hop. *P. falciparum* Hsp70-x (PfHsp70-x) is a malarial protein which has recently been found to be exported outside the parasite during the blood stages of its lifecycle and localise in the host red blood cell cytosol. The first part of this chapter therefore focuses on the potential interactions of PfHsp70-x with other exported malarial J proteins and host J proteins, using both homology modelling and protein-protein docking. Another recent research effort described a structure of a new interaction interface between Hsp90 and Hop in yeast. The second part of the chapter reports structural work of this interaction interface in *P. falciparum*, human, yeast proteins.

The structural characterisation studies involved a great deal of protein homology modelling and remodelling. Chapter 6 was directed at creating a set of Python scripts to automate the various steps involved in homology modelling. As they are scripts that break down the homology modelling process, they may be used to automate as much or as little of this process as desired by the user, as well as quickly test different modelling parameters. This tool has been designed to help users with varying degrees of expertise in homology modelling. It can be used to identify and evaluate templates, as well as perform alignments, modelling and model evaluation. The scripts were assessed and improved through a simple benchmarking test, described in this chapter. By the end of Chapter 6 the structural characterisation of a number of malarial target proteins has been described and potential target sites for drug screening studies were identified. Methods were developed to perform

this type of analysis in other protein targets and scripts were written to aid with this. Additionally, a homology modelling tool was designed to assist with the modelling of proteins for future structural characterisation and *in silico* screening studies.

To tie these two sections of the thesis together, Chapter 7 summarises the results and discusses future work that can be performed. It is explained that this thesis describes the development of an online database containing compounds isolated from South African organisms, and structural work, which includes both the identification of protein target sites to be investigated through ligand docking studies, as well as the development of a tool to model proteins for future studies. These two parts of the thesis both lend themselves to a clear, subsequent step, which is to perform *in silico* ligand docking studies on the protein targets identified in this study. Likewise, the homology modelling tool can be used to either characterise new protein targets or model known targets to be included in these docking studies with the eventual goal of identifying lead compounds, from the South African natural products available, for drug development.

Section 1: Establishment of a South African natural products database

Chapter 1 Introduction to natural products research

1.1 Natural products

Natural products (NPs) are chemical entities that are produced by living organisms [1]. Of specific interest are those known as secondary metabolites, as opposed to primary metabolites which are common organic molecules, such as DNA/RNA, amino acids and some sugars, found in most cells. Unlike primary metabolites, secondary metabolites can perform functions outside a specific cell or organism in which they are produced. This usually involves interactions with other organisms, and is what makes them so useful. These are products of selective evolutionary processes by which they are naturally 'designed' to interact with other biological macromolecules. As such they display exceptional structural diversity and biological specificity when compared to purely synthetic compounds [2, 3]. A good example of this occurs with respect to multiple drug resistance (MDR) pumps, found in certain microorganisms [4]. These expel unrecognised molecular and chemical structures from cells as a defence mechanism. NPs resemble naturally occurring chemical compounds within organisms and making them less susceptible to expulsion by these MDR pumps [5]. A study by Nascimento *et al*. [6] tested the ability plant extracts to inhibit the growth of multiple drug resistant bacteria. Most notable were the clove and jambolan extracts, which displayed antimicrobial activity against 83% of strains tested. It is believed that up to 28% of higher plant species have associated medicinal properties and that nearly 75% of drugs that have been derived from plant extracts were discovered as a result of traditional knowledge [5].

1.2 Traditional medicine

For thousands of years, mankind has turned to nature to treat a multitude of different ailments [7, 8]. This has been witnessed in civilisations from across the world and is referred to as traditional medicine (TM) or complimentary/alternative medicine (CAM). Although there is a distinction between these two they are often used interchangeably. According to the World Health Organization (WHO), the definition of TM includes skills, knowledge and practices used in health care, performed based on experience or belief of different cultures [9]. Practices used in CAM are similar to those of TM, but are not based on long-standing indigenous knowledge, nor do they form part of any official health care system. Globally the demand for TM health care options has increased in recent years as countries have recognised its value [9]. In many rural areas, people have severely restricted access to established health care or qualified physicians and as a result, TM is their only option [10]. TMs have been developed to treat the patient, not necessarily just the disease [11, 12]. This makes treatment more personal and is often preferred for this reason. TM is also considered part of some cultures and although modern treatments are not necessarily outright rejected, they are considered to only be suitable after TM fails [11].

1.2.1 Traditional medicine in Africa

It is estimated that in some African countries, up to 80% of the population are reliant on TM [13, 14]. The use of TM in Africa has been previously frowned upon, as practitioners were associated with witchcraft, which may have contributed to the breakdown in communication between modern medicines and TMs in Africa [10]. In Africa, efforts driven by the WHO were made to raise awareness of TM in countries across the continent and incorporate TM into their health care systems [15]. In 2010 it was noted that 22 African countries, including South Africa, had established studies to test TMs against malaria, HIV, sickle-cell anaemia,

diabetes and hypertension. Mali, where TM is part of the health care system, has established the Department of Traditional Medicine (DMT). This collaborates with the World Health Organisation (WHO), in order to study TM and aims to develop it to be used in tandem with modern medicines [10]. In South Africa, only Western medicines are officially acknowledged [16]; however, it is estimated that more than 3 million people in this country currently use indigenous plants for medicinal purposes [17]. Some medical aid schemes have even incorporated TMs into their cover plans [18]. The South African government and the National Research Foundation (NRF) have made efforts to increase the level of research into South African TMs and natural products. A similar trend has been witnessed on a global scale over the past 30 years [10].

1.2.2 Traditional Chinese medicine

Traditional Chinese medicine (TCM) is probably the most trusted form of traditional medical treatment and is often incorporated into Western medicines [19]. Having been used for thousands of years, this is a very complex form of treatment, but focuses largely on balancing yin and yang within the human body [12]. In additional to treating illnesses, TCM tries to aid recovery of the body, often focusing on the main organs; the liver, heart, spleen, lung, kidney, referred to as ‘organ networks’.

1.2.3 Undesirable effects of traditional medicine

TM treatments for specific ailments may have side effects, although a perceived advantage of TM is its milder nature. When compared to Western medicines, the healing effects of TMs may be less-potent, but so are the side-effects [20]. The side-effects of TMs can, however, be dangerous if administered incorrectly, or if given in combination with certain drugs. A large number of these effects are untested and therefore unknown [9, 13]. Many plants used in TM also produce a number of metabolites that are toxic when ingested [21]. For example, Steenkamp & Gouws [22] reported incidences of patients with severe liver damage, believed

to be caused by pyrrolizidine alkaloids found in TMs they had taken. It was also revealed that this information was not reported back to the traditional healers [23].

1.3 NPs in drug discovery

The discovery and isolation of many medicinal compounds from natural sources can be attributed to known properties of the organisms that produce of these chemical agents. One of the earliest described isolation of a pure compound for medicinal use was that of morphine in approximately 1804, by Friedrich Serturner from opium [24]. At the time, opium was known as an addictive narcotic, though also used in pain management as early as the Middle Ages. The antimalarial drug, quinine was isolated in 1820 by Pelletier and Caventou from the bark of the cinchona tree [25, 26]. This tree bark had been used to treat malaria since the 1600s and also formed part of South American traditional medicine for the treatment of fevers [7]. Another naturally derived analgesic, and one of the most widely used drugs of all time [27], is acetylsalicylic acid, registered under the name "Aspirin" [28]. The precursor to this was salicylic acid from willow (*Spirea ulmania*) bark. The leaves of this plant were used for both their analgesic properties and anti-inflammatory effects by the Ancient Egyptians [28]. The accidental discovery of penicillin in 1928 is relatively well-known. Alexander Fleming is credited with having witnessed that *Staphylococcus* bacterial colonies were unable to grown around a small area surrounding mould of the *Penicillium* genus [29]. Though not derived through known TM, the isolation of penicillin only came through this observation of the organism that produced it. Fleming tested this effect on different microorganisms and deduced that because some were unable to grow even within several centimetres of the mould that it must be producing an antibacterial substance [29]. An increase in drug resistance to quinine and its derivatives led to a need for alternative antimalarials and the isolation of artemisinin from *Artemisia annua* [30]. In TCM this plant has been used for over a thousand

years to treat fevers [30]. The compounds, 5'-methoxyhydnocarpin from *Berberis fremontii*, used in Native American TM, has been shown to inhibit the function of the MDR pump, found in multiple drug resistant bacteria [31]. This indicates a role for NPs in combination with other drugs.

1.3.1 Pharmaceutical potential of marine NPs

Marine NPs are a relatively new source of pharmacological agents [32], but display diverse and unique chemistry [33]. In addition to the drug potential of these compounds, as they occur naturally, they may also be used as template molecules for the design of even more advanced, novel medicinal compounds. In 2014, eight marine NP-derived drugs had been approved, three of which required no additional modification after extraction [34]. These include uses as anticancer agents, antivirals, pain medication and treatment of cardiovascular disease. Between 1951 and 1960, experiments conducted using arabinose-containing nucleosides (spongonucleosides), extracted from the sponge, *Cryptotethia crypta*, found that these had an inhibitory effect on DNA/RNA metabolism [35]. Two of these spongonucleosides, spongothymidine and spongouridine, were modified to produce the drugs cytarabine (anticancer) and vidarabine (antiviral), respectively [34]. Although cytarabine was approved over 40 years ago, it is still a leading modern anticancer drug. The other antiviral is sold over the counter as a nasal spray, Carragelose®, which contains Iota-carrageenan from *Rhodophyceae* seaweeds [34]. The peptide toxin, zinconotide, was isolated from venom produced by the marine snail, *Conus magus* [36]. This required no further modification for the treatment of chronic pain and was approved in 2004/2005 for use in patients suffering from cancer, AIDS and neuropathy [34]. A combination of omega-3 fatty acids from fish oils have been have been artificially synthesized and formulated into the drug, Lovaza®, by GlaxoSmithKline (GSK). This is used to treat severe hypertriglyceridemia, based on the observation that individuals native to Alaska, whose diets were high in these omega-3 fatty

acids, were far less prone to cardiovascular disorders [37]. The compound, trabectedin, was first isolated from the tunicate *Ecteinascidia turbinate* [38]. It displays potent anticancer activity and is currently marketed as the drug, Yondelis® [34]. The yield of this compound from its source organism is too low to be viably extracted for drug production, but can be semi-synthesised using a much higher-yielding NP, cyanosafracin B, from the *Pseudomonas fluorescens* [39]. Another anticancer compound, halichondrin B, was isolated from the sponge *Halichondria okadai* in 1986 [40]. Again, due to low yields, the compound needed to be synthesised to be studied. This allowed the derivative compound, eribulin mesylate, to be discovered, which is currently approved as an anticancer drug [34]. The final anticancer marine drug is brentuximab vedotin, which is a structural analogue of dolastatin 10 from sea hare *Dolabella auricularia* [41].

1.3.2 Recurrence of NPs in drug discovery

The process of producing a novel, approved drug is very long and incredibly expensive [42]. In the 1990s, high throughput screening (HTS) methods were introduced in order to speed up drug discovery process. This allowed companies to screen large synthetic chemical libraries to test for specific biological activity [42]. NPs are not suited to HTS and many pharmaceutical companies consequently discontinued their NP drug discovery programs [43]. The decline of NPs in pharmaceutical drug discovery programs has been attributed to a number of factors. From the side of industry, these included 1) lack of technologies available in the 1990s for NP-based drug discovery; 2) pharmaceutical companies trying to investigate too many different sources at the same time, rather than giving sufficient attention a small number of NPs; and 3) a lack of expertise in industry for establishing and maintaining NP libraries [44]. NPs themselves are also inherently more difficult to work with than synthetic compounds. They are generally produced in small amounts, as part of a mixture with other small molecules and biomass. This makes obtaining sufficient quantities of pure compounds

for testing labour-intensive, time-consuming and ultimately highly costly [44]. The process of NP screening is also prone to re-discovery of an already identified NP with specific activity. This is because some NPs are isolated more frequently than others and dereplication methodologies don not always identify these as the same compound [44]. Finally, the structural complexity of NPs makes them difficult to work with, modify and synthesise [44]. Recent advances in technology and chemical methodologies has seen some of these challenges addressed and NPs are being reintroduced into the drug discovery process [2]. This has also been due to the failure of HTS technologies to produce the results expected. HTS libraries use combinatorial chemistry to produce a myriad of new chemical entities (NCEs). Unfortunately, these technologies identify compounds with specific biological activity and do not distinguish whether or not they will be a good or safe drug component [42]. HTS also has low selectivity, as approximately 50 - 85% of tested compounds will go through to subsequent stages in the drug discovery process [45]. Even so, the process of combinatorial chemistry with HTS has only yielded a single NCE that has been approved as a drug [2, 8].

1.4 NPs of South Africa

South Africa ranks third in the world in terms of its terrestrial biodiversity [46], and it is estimated that 3000 different plant species are used for medicinal purposes in this country, mostly as part of traditional medicines [47]. It is estimated that more than 3 million people in this country currently use indigenous plants for medicinal purposes [17]. Organic extracts from South African plants have been shown to display both antibacterial and antifungal activity [48, 49]. Additionally, chemical compounds extracted from these plants have shown promise as agents against cancer [50], *Mycobacterium tuberculosis* [51] and various neurological disorders [52]. Clarkson *et al*. [53] reported at least moderate antiplasmodial

activity from 49% of the 134 different South African medicinal plant extracts tested. Likewise, Bessong *et al*. [54] reported a series of South African plant extracts that displayed anti-HIV activity. The antimicrobial activity of South African plant life is reviewed by van Vuuren [55], including examples of specific compounds isolated from these plants and their measured antimicrobial activity. Kelmanson *et al*. [18] reported screening of 14 different plants used by Zulu traditional healers and found a wide variety of antibacterial activity, mostly against Gram-positive bacteria. Some women in South Africa use TMs to help with labour and antenatal care. Kaido *et al*. [56], tested some of these treatments and found them to regulate acetylcholine and oxytocin.

Although early research attempts were made, Southern African marine chemistry research only really took off in the 1990s with compounds isolated showing a great deal of potential as anticancer agents [57]. Further, South African marine organisms have been shown to synthesise many novel compounds which also display pharmaceutical potential [58, 59]. The marine and plant life of South Africa represents an untapped resource of chemical compounds.

1.5 Research motivation

NP scaffolds are an important component in the drug discovery process [3]. With regard to the chemical structure of morphine, Rishton [60] noted that it is one that would never be achieved through rational drug design by medicinal chemists, thus highlighting the value of its discovery from a natural source. Additionally, nearly 65% of all drugs approved between 1981 and 2010 were either derived from or inspired by NP precursors [61]. More recent studies reveal that since 2008 25 new NP-derived drugs were approved, and in 2014 there were 31 additional drugs either at or past phase III clinical trials [62]. The development of *in silico* methods for drug discovery has presented faster and cheaper ways to screen for

compounds with pharmaceutical potential [63]. This approach is meant to supplement and direct *in vivo* screening, which is required for experimental proof that a drug is effective. With on-going development in this area of research, *in silico* drug screening methods are becoming more accurate and more reliable with time [64]. Online databases, such as ZINC make compounds freely available for *in silico* screening [65]. For similar reasons, a number of NP databases have been developed, though are mostly dominated by compounds extracted from traditional Chinese herbs and medicines. The flourishing NP research in Africa has also resulted in a number of NP databases [66–68] emerging in recent years, though nowhere near the scale of their Chinese counterparts. None of these African databases have an online interface to access this data, nor do they focus on NPs from South Africa, where there is a wealth of chemical biodiversity. Data concerning potential medicinal compounds from organisms indigenous to this country has become severely scattered across a variety of databases, as well as hundreds of journal articles which often require a fee to gain access to [69]. Additionally, information in literature is not easily searchable, especially when attempting to acquire specific information or finding compounds with specific properties. It would be greatly beneficial to South Africa if compounds from this region were freely available and easily locatable and accessible for research purposes. This would increase exposure to these small molecules and enhance the range of compounds tested as inhibitors against potential drug targets.

1.6 Research aim

The focus of this project was to make a resource available containing compounds from South African organisms. As such, the overall aim is to develop a web-based database containing biologically relevant information about compounds extracted from South African plant and marine life. This is divided into the following objectives:

1. Design and create a database to store information regarding South African natural products
2. Populate the database with relevant information
3. Establish a web interface which would enable anyone to access this information for their own research purposes.

Chapter 2 South African natural products database

There are a myriad of well-established databases around the world, which specialise in information pertaining to chemical compounds, both natural and synthetic. This chapter reviews a number of these databases and their uses, as well as the ways to work with chemical data. Described herein is the design and construction of a database of South African natural products, which aims to contain fully-referenced information from literature regarding these compounds, the organisms from which they were isolated and their uses. Information about each compound was extracted manually from literature and linked to external databases, where possible. The physicochemical properties of all compounds were calculated and each compound was converted to different file formats to allow them to be used in computational experiments. The uploading of compound data was subjected to manual curation and various error-checking steps to ensure data quality. The current content of the database is described, as well as potential plans for the future developments that can be made to enhance its value. The work presented in this chapter and Chapter 3 was recently accepted for publication [70].

2.1 Introduction

2.1.1 Current compound databases

Compound databases exist to describe and quantify our knowledge of chemical entities and assist with the identification of new entities [71]. Unlike other databases, they may contain information about the structures of chemical compounds, as well as reaction data. To date, a number of chemical compound databases have been created, each with a specific purpose. The most well-established databases, described below, contain millions of records with in-depth information about each compound.

2.1.1.1 PubChem

PubChem is an open online repository consisting of three related databases, namely the Compound, Substance and BioAssay databases [72]. It was developed by the National Center for Biotechnology Information (NCBI) as part of the US National Institutes of Health (NIH) program to establish of molecular libraries [73]. Together these contain information about small molecules and their biological properties. The Substance database contains records of all deposited substances and the BioAssay database contains information regarding screening studies against gene and protein targets. The Compound database contains unique entries from the Substance database, after processing and standardising the data. In 2010, PubChem contained over 25 million compounds and 90 million bioactivity data for thousands of protein targets. The database was developed to help researchers with drug development [74].

2.1.1.2 ChEMBL

ChEMBL was developed at the European Molecular Biology Laboratory (EMBL). This is a curated bioactivity database, containing information manually extracted from literature. The data contained within ChEMBL includes binding activity, functional bioassay data, and absorption, distribution, metabolism, excretion and toxicity (ADMET) properties for over a million compounds. The only compounds considered are those with drug-like properties [75].

2.1.1.3 DrugBank

DrugBank contains a wealth of information about approved and experimental drugs and drug targets [69], including definitions of drug molecules as well as detailed information about their chemical and pharmaceutical properties. The database is ever-expanding to include additional information about each drug, such as detailed drug pathways and interaction data in DrugBank 3.0 [76] and recently more detailed ADMET profiles and quantitative structure activity relationships (QSAR) information with version 4.0 [77].

2.1.1.4 ZINC

The ZINC database was started with the goal of providing a set of quality compounds that could be used for *in silico* docking studies. The database also contains information pertaining to the purchasing of molecules from different chemical vendors. ZINC contains approximately 20 million compounds, which can be downloaded in 'ready-to-dock' format. Searches can be based on biological activity, structural properties, as well as an assortment of different chemical identifiers. This is a rapidly-expanding database, growing nearly more than 10-fold in size since it was started in 2005 [65, 78].

2.1.1.5 Other notable databases

A number of other well-established databases have also been developed. A few of these will be mentioned to give an indication of the variety of information collected about different compounds. BindingDB [79] contains experimentally-determined information about protein-ligand binding affinity, obtained through literature searches from scientific publications. It was constructed to allow the study of protein-ligand interactions in order to determine the ideal chemical properties for pharmacophores, gain mechanistic insight into specific ligand binding and provide a means to check for side effects of potential drugs. The Cambridge Structural Database (CSD) is a resource at the Cambridge Crystallographic Data Centre (CCDS), containing records of small molecules solved by crystallography (both X-ray and neutron) [80]. It currently has records for over 700000 chemical structures, with nearly 40000 added on a yearly basis. Chemical Entities of Biological Interest (ChEBI), developed at the European Bioinformatics Institute (EBI), is described as a dictionary for small chemical compounds, containing information about compound ontology [81]. This describes the relationship between a small molecule and other molecules directly associated with it. ChEBI aims to provide the "most correct" nomenclature for molecules, as specified by relevant authorities (e.g. IUPAC for names of chemical compounds), as well as ensure that all items in

the database are fully referenced, so that users can find the source of specific information. A number of metabolome databases, such as Biospider [82], The Human Metabolome Database (HMDB) [83] and the Yeast Metabolome Database (YMDB) [84] detail molecules with respect to the annotation of metabolomes. A similar database, the Small Molecule Pathway Database, describes over 350 small molecule pathways found in humans, which was designed to help with clinical metabolomics and related research [85]. Finally, the Toxin and Toxin-Target Database [86] contains biologically important data about toxins, including mechanisms, chemical structures and clinical data. This includes over 3000 toxins and over 33000 associations with 1300 targets. It is a highly comprehensive, curated database with up to more than 80 fields of different information for any given toxin or its target.

2.1.1.6 Natural products databases

NPs account for a great deal of known drugs so researchers have started to develop their own chemical libraries to screen NPs for drug potential [87]. Some NP databases have also been developed to further understanding the mechanisms of action of traditional medicines [88]. Since NPs are produced to interact with biological targets [2, 3], the design of these databases provide a biologically- and therefore pharmaceutically-relevant chemical space to work with [89]. The NAPRALERT database [90] was established in 1982 with the endeavour to extract compound information from literature since 1880. The database has been used to aid the development pharmaceuticals and cosmetics. Unfortunately this is a commercial resource, limiting its use in academic research. More recently, there have been a number of freely accessible NP databases established as discussed below.

2.1.1.6.1 Traditional Chinese medicine databases

Traditional Chinese medicine (TCM) is an incredibly well-established source of natural products data [91]. Some of the most well-established TCM databases include the TCM Database@Taiwan [91], CHDD [92] and the Chinese Natural Product Database (CNPD)

[93], which together have nearly 80,000 unique chemical structures [94]. These databases contain compounds found in TCM and information is collected from Chinese medical texts, as well as scientific literature. The TCM Database@Taiwan was designed in response to a great deal of interest being directed towards TCM for potential lead compounds in drug design, especially by pharmaceutical companies. These compounds are prepared in such a way that they should be readily usable for *in silico* drug screening [91].

2.1.1.6.2 Natural products database of Brazil

The NuBBE database ($NuBBE_{DB}$) was established to catalogue compounds isolated from the biodiversity of Brazil [95]. Since the database was established only fairly recently, it contains only 640 compounds from scientific publications. It has been integrated into a web-based interface, designed to help with natural products research.

2.1.1.6.3 Natural products databases in Africa

Currently, in Africa there is only one group with an official NP database. This was originally established to contain compounds isolated in Cameroon and was named CamMedNP [66]. This later expanded to include compounds isolated from the entire Congo Basin and therefore named ConMedNP [67]. The Congo Basin is rich in diverse flora and plant life in this region is a favoured medicinal source. The database currently contains just short of 3200 compounds and largely focuses on providing 3D structures of compounds, specifically for use in *in silico* drug design. The same group created a similar database, AfroDB, which contains a relatively small number of compounds from across the African continent [68]. The database was kept small, so as to be used in *in silico* screening by groups in Africa with limited computational resources. In order to achieve this, the database only contains compounds with established biological activity, such as anticancer activity [68].

2.1.1.6.4 Marine natural products database

Since most marine organisms rely on purely chemical defence mechanisms they need to produce potent compounds with great biological specificity [2]. Lei & Zhou [96] have developed a database containing over 6000 marine natural compounds isolated since 1960. The database is predominantly for studying marine chemical entities to discover new pharmaceuticals; however, it does also provide benefits to research of marine toxins and chemical ecology.

2.1.2 Working with chemical data

Line-formula notation, developed in 1861, was used to describe chemical structures by separating consecutive atoms within the main chain of a compound by periods ('.') while other adjacent atoms were written before the next period [97]. A number of research groups used their own variations of line-formula notation to further develop it [97], but the convention was considered difficult to work with for those without extensive knowledge of it [98, 99]. Since the involvement of computer systems in handling chemical data, it became important to use standardised notations for representing chemical structures that computers can recognise and work with [98]. The most widely used formats still in use include SMILES, CAS, molecule file (Molfile) and structure-data file (SDfile) [98, 100].

2.1.2.1 SMILES

Simplified Molecular Identification and Line Entry System (SMILES) notation was developed as an alternative to line-formula notation, which could be used by computer scientists with a limited background in chemistry [99]. It allows a compound to be written as a text string and follows five rules. 1) Each atom is represented by its atomic symbol. If an atom has a two-character symbol (e.g. Cl), then the first character is written in upper case and the second in lower case. Hydrogens do not need to be written in, as they are implicit, but can be included if required to avoid ambiguity; 2) Bonds between atoms are shown as dash

characters (‘-’) for single bonds, equal sign characters (‘=’) for double bonds and a hash characters (‘#’) for triple bonds. If a bond is not specified, it is considered to be a single bond. 3) If atoms branch off the main chain of the chemical structure, these are enclosed in parentheses; 4) Cyclic structures (rings) are denoted with a digit input immediately after the atom that opens a ring and the same digit immediately after the atom that closes that ring; 5) If a ring is aromatic, then the atoms are written in lower case. If an elements in a ring has a two-letter symbol an exclamation mark (‘!’) is written immediately after the atom, to avoid ambiguity [99, 101]. Examples of chemical structures with their SMILES are shown in Figure 2.1. Additional SMILES specifications can be used for configuration around a double bond and tetrahedral centres [102]. Chirality is specified using the “at” character (‘@’) in square brackets, as shown in Figure 2.2 (A). If a single character is used, the branched atoms follow in an anticlockwise order (left), whereas two characters indicate that these follow in a clockwise order (right). Double bond configuration is specified using back and forward slash characters, as shown in Figure 2.2 (B). If two of the same slash characters are used for double bond atoms, it represents *trans* (*E*) configuration (left), whereas if both forward and back slashes are used, it represents *cis* (*Z*) configuration (right). Square brackets can also be used to specify charged atoms, such as [NH+] [102].

2.1.2.2 SciFinder

SciFinder is a commercial research tool, developed by the Chemical Abstracts Service (CAS), which consists of a set of five databases containing information about chemical literature, substances, reactions and suppliers [103]. Over 93 million compounds are contained within the CAS registry [104], each with a unique CAS Registry Number (CASRN). The CASRN is a universal and authoritative identifier, used by chemical vendors and regulatory agencies [71].

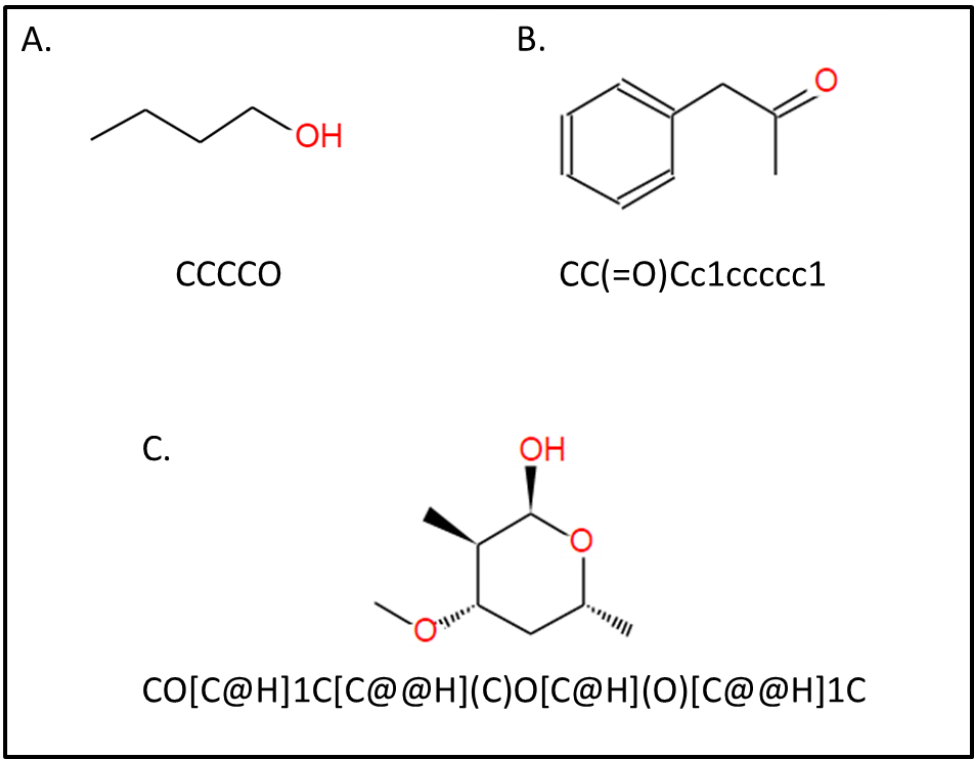


Figure 2.1: Examples of compounds and their associated SMILES. Compounds were drawn and the SMILES produced using the JSME molecular editor applet [105]. Structure (A) represents a simple aliphatic compound; (B) shows a an aromatic ring, as well as the use of displaying branches in SMILES, and (C) shows an example of a non-aromatic ring with the stereochemistry fully defined.

2.1.2.3 InChI and InChIKey

IUPAC Chemical Identifier (InChI) was designed to be a non-proprietary, unique identifier for a chemical structure [98]. It is described as being a type of bar code that also contains structural information, meaning this data does not need to be retrieved from a database, as is the case with the CAS identifier. InChI is a layered representation with each layer, separated by a forward slash, describing different types of information. InChI can contain a number of different layers that together specify the formula, connectivity, isotopes, stereochemistry and tautomers of a chemical structure. Since the length of an InChI string will increase with the size of a compound, an InChIKey has also been developed. This is a 27-character representation of an InChI, which can be used more easily to find a chemical structure either in a database or an internet search engine [98].

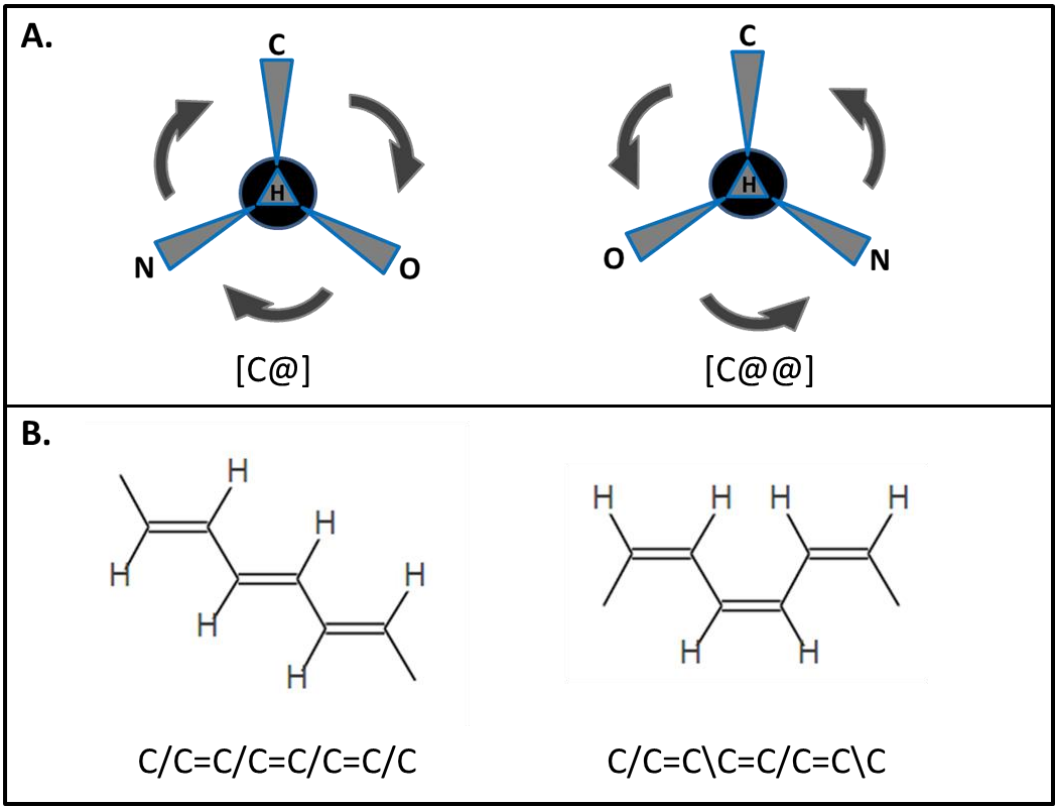


Figure 2.2: Examples of stereochemistry specifications made using SMILES. Structural representations are shown which indicate the different ways to specify stereochemistry when writing SMILES. The SMILES is shown below each structure. (A) displays the use of the '@' symbol to represent the chirality of a tetrahedral centre. The black circle represents a tetrahedral carbon and each triangle represents a single bond to either a Carbon, Hydrogen, Nitrogen or Oxygen atom. (B) displays the use of forward and back slash characters to specify the configuration around a double bond. On the left, the slashes all face the same way, representing *trans* configuration, whereas on the right they face opposite directions, indicating *cis* configuration.

2.1.2.4 Chemical table files

Another series of formats that have been developed to allow computers to use chemical structural data are those based on the chemical table file format, designed by Molecular Design Limited [100]. Of these, Molfile and SDfile are the two used by major databases such as ZINC, ChEMBL and PubChem. The Molfile consists of a header block and a connection table. The header block contains information about the molecule, who created the file and other miscellaneous data. The connection table is common to all chemical table file formats. It is divided into six different blocks, which together describe the structural relationship between all atoms in a molecule, as well as their properties. The SDfile is an extension of the

Molfile and can hold structural information for multiple compounds. Each compound entry in an SDfile has a Molfile section, containing the same information as this format. Additional data sections are included, which provide more information, for example the melting point of the compound. Each data section contains a header, indicated by a greater than sign ('>'), with the associated information on the following line and separated from other data sections by a blank line. Different compounds in an SDfile are separated by four dollar signs ('$$$$').

2.1.3 Proposed work

This chapter will focus on describing the design and layout of a South African NP database, including the collection of compound data from literature, as well as other information considered relevant for further study of these compounds. Additionally, the current content of the database is assessed, along with means of quality control of data. Future plans for further development of the database are also described.

2.2 Methodology

2.2.1 Database

SANCDB comprises a MySQL database, integrated into a Django application. The database schema is shown in Figure 2.3. The database is centred around the Compounds and References tables. The References table was linked to the Authors table, via the RefAuthors table, which, apart from linking these two tables, also keeps track of the first author of a publication. The References table is also linked to Journals and Depositors, each as a lookup table. The Depositors table houses information about individuals who upload information to the database. The Compounds table contains information to identify a compound (name, formula etc.), as well as identifiers that link them to external databases, such as PubChem. Compounds are linked to the References, Sources, OtherNames and Uses tables. Sources refer to the organisms from which a given compound was isolated and Uses refer to any specific activity tested for a compound; e.g. antimalarial activity. OtherNames entries are captured as the name used to refer to a compound in a specific publication, as well as names assigned to the compound by SciFinder [106] if applicable.

2.2.2 Django

2.2.2.1 Models

The tables from the database layout, presented in Figure 2.3, were written as a set of Python [107] classes, referred to as 'models' in Django. These are used by Django [108] to create a MySQL [109] database containing the information specified in the schema and allow the database to be manipulated as python objects. These are specified in the file, **models.py** (refer to Appendix A).

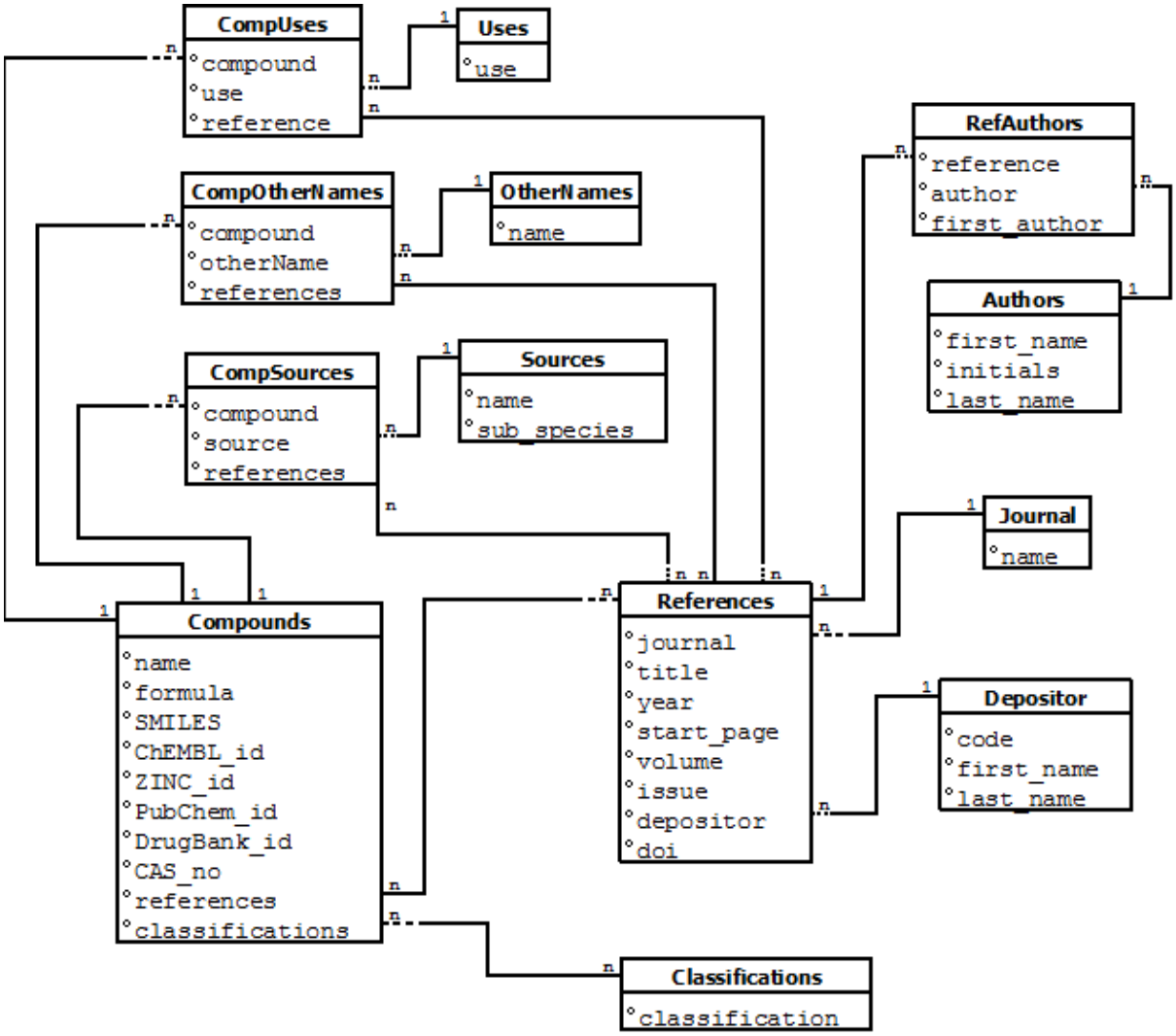


Figure 2.3: ER Diagram for the SANCDB, describing the layout of the database and the relationship of the data. Each container represents a table in the database. The table name is shown at the top of each table with the different fields listed below. Links between the tables represent the relationships between the data as either a 1 (one) or n (many). Table and field names displayed are those used in the database.

2.2.2.2 The admin site

Django provides an admin site that creates an interface which allows easier control and visualisation of data, especially by annotators with little general scripting experience. Each model was registered with the admin site. This involves specifying which fields in the models are displayed, how they are ordered and which can be searched. These are specified in the file **admin.py** (refer to Appendix A).

2.2.3 Literature search

Publications, theses and textbooks were searched for, focusing on work involving the extraction of compounds from South African plant and marine life. The Rhodes University thesis collection [110] was searched for theses by students that match these criteria. PubMed [111] and ScienceDirect [112] were also searched for publications matching these criteria. Additional literature was found using Google searches [113] and searching for publications and textbooks referenced in literature already found.

2.2.4 Compound uploading

2.2.4.1 Admin site

The Django admin site was used to upload compounds. Depositors were registered as users and given permission to upload and edit data in the database. Information was uploaded in a specific order. Reference information was always uploaded first and a depositor was linked to the reference. When compound information was uploaded, the reference entry was linked to the compound. All information, apart from SMILES (refer to Section 2.2.4.2), classifications (refer to Section 2.2.4.3) and external links, was obtained from literature. Each compound was searched for in SciFinder and a CASRN was assigned to it, if available. This was also used to ensure the compound uploaded was not already in the database. The Sources, Uses and OtherNames tables were also populated using information from publications and linked to both the Compounds table and the References table via the CompSources, CompUses and CompOtherNames tables, respectively (Figure 2.3).

2.2.4.2 Compound SMILES generation

SMILES information is not readily available in literature, but an image of each compound is usually provided. To ensure that all known structural information of the compounds was captured, images were downloaded from SciFinder for compound entries. These were

uploaded to the OSRA compound image recognition server web interface [114], which reads a compound image and draws out the structure in the JME molecular editor [105]. This molecular editor can be used to correct mistakes made by OSRA and produce SMILES for the correct structure. The Daylight depict server [115] was used to check the structure produced by the SMILES and Open Babel [116] was used to produce InChI for the SMILES. This was used to ensure the SMILES matched the formula of the compound and make sure all known stereocentres were specified.

2.2.4.3 Compound classifications

Classifications for each compound were assigned by Dr Kevin Lobb from the Rhodes University Chemistry Department.

2.2.5 Compound image generation

Compound image generation was performed using the Indigo tool kit [117]. This is a Python module that requires compound SMILES as input. A Python script was written which uses this module to automate the generation of compound images, by querying the database for the SMILES of each compound (Appendix B; File: **comp_imgGen_indigo.py**).

2.2.6 File format generation

Compound 3D coordinates were calculated from SMILES using the CORINA structure generator [118, 119], saved in SDfile format. Open Babel [116] was used to convert the 3D coordinates to Molfile and PDB format. Open Babel was also used to convert each compound PDB file to GAMESS input file format. GAMESS [120] was then used to minimize the compounds, using RM1 basis level. These were then converted back to PDB, using Open Babel and saved separately as ‘minimised’ PDB files. A Python script was written to automate this process (Appendix B; File: **CORINA_conversions.py**).

2.2.7 Calculation of compound properties

Properties were calculated for each compound, based on Lipinski's rule of five [121], which requires the molecular mass of the compound, the calculated logP (cLogP), as well as the number of hydrogen bond donors (HBDs) and hydrogen bond acceptors (HBAs). Molecular mass was calculated from the compound's formula, using known average mass of each element. The obprop program in the Open Babel installation was used to calculate cLogP values. A SMILES parser has also been developed (refer to Section 2.2.9.4) and was used to calculate both the HBAs and HBDs present in each compound. A Python script was written to automate this process and parse the output from Open Babel, as well as to save the values recorded (Appendix B; File: **memcache_parse_babel_props.py**).

2.2.8 Validation and backup

SMILES produced by OSRA were checked primarily by calculating the formula of the compound from the SMILES produced, using Open Babel. This was compared to the known formula of the compound from literature. If these did not match, it indicated that the compound was incorrectly drawn by OSRA and not correctly fixed by the depositor before saving the SMILES. External IDs were checked by their format, e.g. DrugBank compound IDs start with the letters 'DB' and are seven characters in length. Spot checks were performed by looking at random entries in the database and comparing the values captured to those found in literature to ensure information was accurately captured and no information was omitted.

2.2.9 SMILES parser

A SMILES parser was developed as a Python module to read and work with SMILES data. It is split into four different classes. The full script can be viewed in Appendix C; File: **smilesparser.py**.

2.2.9.1 Atom_position class

This class is used to assign a position level to each atom in a SMILES string, which identifies where each atom is located in the compound, based on the way the SMILES is written. Atoms in the main chain are numbered using integers, (i.e. 1, 2, 3, etc.). Branching from the main chain is denoted with a colon (':'), which is also used to denote further branching. Each branch is numbered, as are specific positions on each branch, e.g. atom at position '3:2-5' is the fifth atom along the second branch of the third atom in the main chain. The Atom_position class also contains functions, which enable navigation through a SMILES string, based on these position names.

2.2.9.2 Square_brackets class

The Square_brackets class simply deals with the different types of information that can be contained in square brackets in a SMILES string. It identifies if the atom has a charge ('+' or '-') or a set chirality (either '@' or '@@') and also identifies the atom within the square brackets.

2.2.9.3 Atom_class

The Atom class represents an individual atom within a SMILES string. It contains an element identifier (e.g. 'C' for Carbon), an Atom_position object, ring information, a pre-bond identifier, charge information, chirality and double bond configuration. The ring information is taken directly from the SMILES as the integers that indicate if the atom either starts or ends a ring. The pre-bond identifier is either a dash ('-'), equals sign ('=') or hash ('#'), denoting a single, double or triple bond (refer to section 2.1.2.1). Charge information and chirality are assigned using the Square_brackets class and double bond configuration is saved as either a forward or a back slash if stated in the SMILES.

2.2.9.4 SMILES class

The SMILES class parses a SMILES string and separates in into a list of Atom objects. It also uses this information determine the positions of all rings in the structure (i.e. which atom objects make up each ring).

2.3 Results and discussion

2.3.1 Database content

2.3.1.1 Number of compounds

The database currently contains 600 pure compounds isolated from South African organisms. This information was manually extracted from 170 different references. The closest comparison to this is the NuBBE database in Brazil, which contains 640 compounds from 170 different references [95], although nearly 100 of these compounds (14%) are either synthetic, semi-synthetic or products of biotransformation. Often in publications, researchers perform simple modifications to compounds. For example, Elgorashi *et al*. [122] used acetylation to convert one of the compounds isolated in the study, Lycorine, to 1, 2, Di-O-acetyllycorine. This was achieved by simply stirring the compound with equal volumes of acetic anhydride overnight. These modified compounds have been omitted from the database, as these are not pure natural products; however, in future it may be worth including the information. A well known South African chemist, Frank Warren [47] has published two reviews of pyrrolizidine natural products [123, 124], which mostly contain these kinds of modifications. Efforts are constantly being made to expand the database. Searches for new literature have become more challenging now that so many papers have already been included, especially with keeping the database restricted to South African NPs. Expanding the database to include NPs from Southern Africa would be the next logical step, since this entire region does not have a NP database, with the exception of AfroDB [68], which contains just over 200 compounds from Southern Africa, but does not have a web interface for users to search or download compounds. This is further discussed in Chapter 3.

2.3.1.2 Compound classifications

Compound classifications were initially taken from literature. However; these can be subjective and were not always consistent for the same compound found in different publications. As a result, classifications were assigned by inspection of each compound. This was done by Dr Kevin Lobb of the Rhodes University Chemistry Department.

Classifying chemical compounds is a means of sorting large sets of data into smaller, logical groupings, which are more manageable and easier to work with [125]. The compound classifications used in the database are structural classifications, where a certain class of compound will have a single common structure. More specific classifications can also be assigned within a class of compounds. This structural hierarchy is useful because compounds from a specific classification may have common properties that are of interest to specific research [125]. Some chemical classifications are not strictly defined, making it difficult to assign objectively. ChEBI [81] is a dictionary for small molecules involved in biological processes; either naturally or by artificial intervention. It gives structural classifications of compounds through the use of ontologies, providing a definition of each chemical class at different levels. Hastings *et al*. [125] have described some examples of problems with the way certain compound structural classifications are defined. Firstly, the words "usually" and "not normally" are used in the definitions of alkaloids and steroids, which indicate that assigning some classifications requires a certain level of discretion. Secondly, the level of chemical expertise required to decisively classify a compound into its most specific structural class makes this process challenging. For this reason the classifications assigned in the present South African NP database are suggested classifications. A total of 116 different compound classifications have been assigned and where possible, more than one classification was assigned per compound, to be as specific as possible. The division of compound classifications within the database is presented in Figure 2.4.

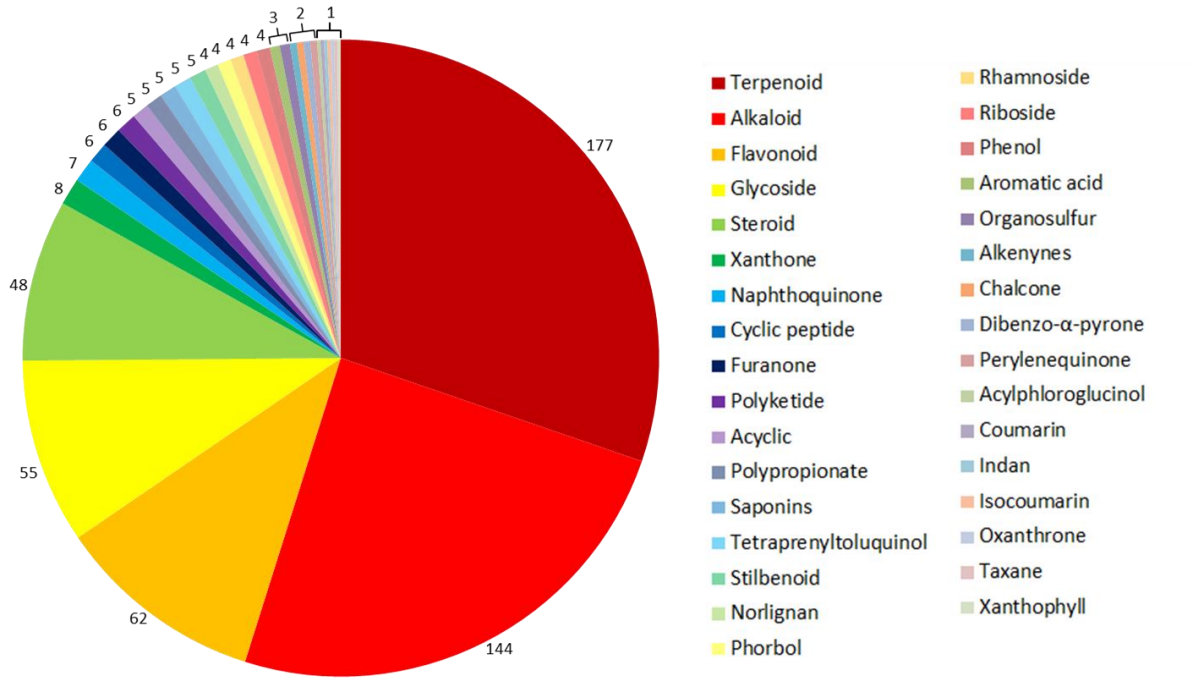


Figure 2.4: Pie chart indicating the distribution of compound classifications within the database. Classifications shown are those which do not fall into a broader classification, as specified within the database. The numbers of compounds that fall into each classification are shown next to each slice in the pie chart. Slices labelled using square brackets each contain that many compounds. Adapted from Hatherley *et al.* (in press) [70]**.**

The main classes of compounds represented in the database include alkaloids, flavonoids, glycosides, steroids and terpenoids, together making up 83% of the assigned compound classifications. This figure is slightly lower, however, as some of these compound classifications overlap (Figure 2.5). Of the 144 alkaloids, 14 are also triterpenes, a sub-classification of terpenoids. There are 19 compounds in the database known as cephalostatins, which are steroidal alkaloids. There are also 10 pregnanes, which are classified as both steroids and glycosides. Figure 2.5 shows how within the five most common compound classifications in the database, there are a number of sub-classifications. Many of these are also further classified to create more specific groupings. This gives a small glimpse into the diversity of compounds in the database.

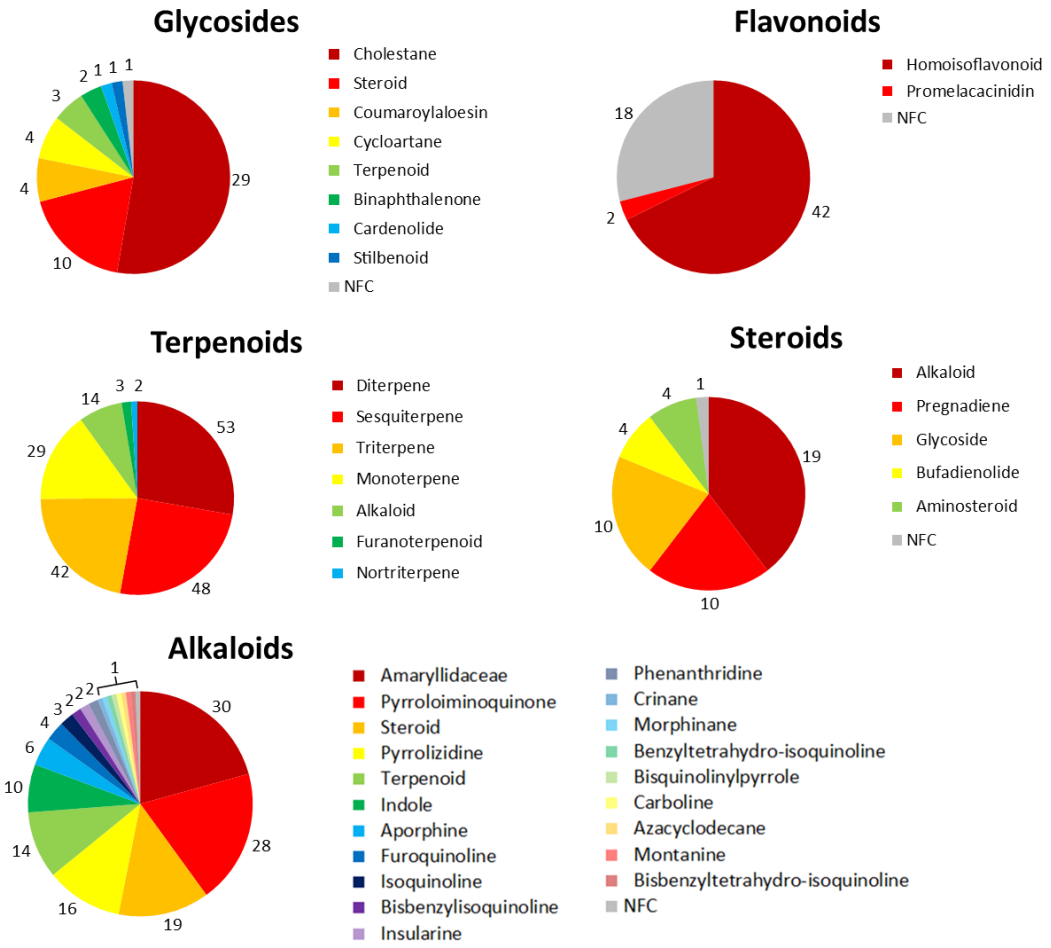


Figure 2.5: Sub-classifications of major compound classifications present in the database. A pie chart is displayed for each of the five most prevalent compound classes in the database, showing more specific groupings of compounds within.

Alkaloids from plants have been used globally in both medicines and poisons; as well as established drugs such as quinine, morphine and caffeine [126]. These are a structurally diverse set of compounds, many of which are considered to have a great deal of potential in modern drug discovery efforts, especially as antibacterial and antiviral agents [126–128]. Flavonoids, as discussed by Cao *et al*. [129] are an incredibly important class of compounds present in foods with health benefits ranging from antibacterial to anticancer properties. Flavonoid compounds have been tested and shown to display antibacterial, antifungal and antiviral activity [130]. Currently there are strategies to use bacteria to modify (biotransform) known flavonoids to create novel molecules with pharmaceutical potential. Mouradov &

Spangenberg [131] explain that the evolution of this class of compounds helped establish terrestrial plants and often form part of chemical defence mechanisms and protection from ultraviolet light. Cardiac glycosides have been used to treat heart problems and, more recently, cancer [132]. Steroid glycosides are common in marine organisms and mostly differ to terrestrial glycosides. These are structurally diverse, but very little has been done to assess their pharmaceutical potential [133]. Most steroids that have important pharmaceutical applications are made by modification of NPs, rather than synthesising them directly [134]. Terpenoids are the largest class of NPs [135] and make up the largest portion of the current database. This class of compounds is associated with a number of natural chemical defence processes and biological interactions, and have a number of well-characterised pharmaceutical properties [135, 136].

When comparing the collection of South African compounds collected in the present work to those from other NP databases, only CamMedNP and $NUBBE_{DB}$ [66, 95] give an indication of the different compound classifications present in their databases. As with our collection, terpenoids were the most common classification in both databases. This makes sense as terpenoids are the largest class of natural products [135]. Flavonoids and alkaloids were also prevalent in both databases, though to a relatively lesser extent for alkaloids in CamMedNP. The alkaloid-richness of our database may prove significant as this group of compounds are relatively underrepresented in pharmaceutical trials, even amongst NPs, despite their past successes, and it is believed that alkaloids should be used in drug discovery efforts [126–128]. Our dataset also contains a large portion of glycosides and steroids, which make up a small percentage of CamMedNP and are not mentioned in $NUBBE_{DB}$. There is no mention of some of the lesser-represented chemical classes from our database, such as organosulfurs, indans and taxanes in either database. These do, however, make up a small portion of the South African compounds, so it is possible that these classes fall into the "other"

classifications section mentioned in both CamMedNP and $NUBBE_{DB}$. For CamMedNP, this would be challenging to confirm. The lack of a web interface prevents compounds to be searched by their classifications. The $NUBBE_{DB}$ website does allow compounds to be searched by classification (listed as 'Chemical Class'), but provides no indication of what classifications are present in the database. No compounds were returned when searching the terms 'acyclic', 'acylphloroglucinol', 'alkenyne', 'aromatic acid', 'furanone', 'indan', 'norlignan', 'organosulfur', 'oxanthrone', 'perylenequinone', 'phorbol', 'polypropionate', 'riboside', 'stilbenoid', 'taxane', 'tetraprenyltoluquinol' and 'xanthophyll'. Due to the partially subjective nature of classifying compounds, it is possible that some of these classes of compounds are present in $NUBBE_{DB}$, but were simply assigned different classifications.

In future, each compound in the database needs to be classified down to its most specific structural class, in order to make filtering compounds by classification as useful as possible. Unfortunately, this is not a trivial task. Apart from the problems already mentioned, compound classifications are not always logical from a structural point of view. A good example of this is highlighted by Bobach *et al*. [137] when they compare two of the compounds shown in Figure 2.6, androstane (A-1) and pregnane (B-1). Both are considered to be steroids, but fall under different classifications, which holds true for their derivatives. Even though the androstane skeleton is a common substructure in all compounds in Figure 2.6, only abiraterone (A-2) is considered to be an androstane derivative. The compound pregnan-21-al (B-2) is a pregnane derivative and not classified as an androstane. This distinction between the two classifications is linked to the change in biological activity caused by the small structural difference [137]. As such, assigning classifications involves not only finding a suitable substructure, but also ensuring that a better substructure does not exist, which accounts for modifications, as with androstane vs pregnane derivatives. To do this manually would be both challenging and time consuming. Strategies are being developed

to automate this process to some extent by using information from chemical ontologies, such as ChEBI, with some success [125, 137]. This may need to be the next step taken with compounds in this database, as this would provide consistency and information would have set definitions from an authoritative source.

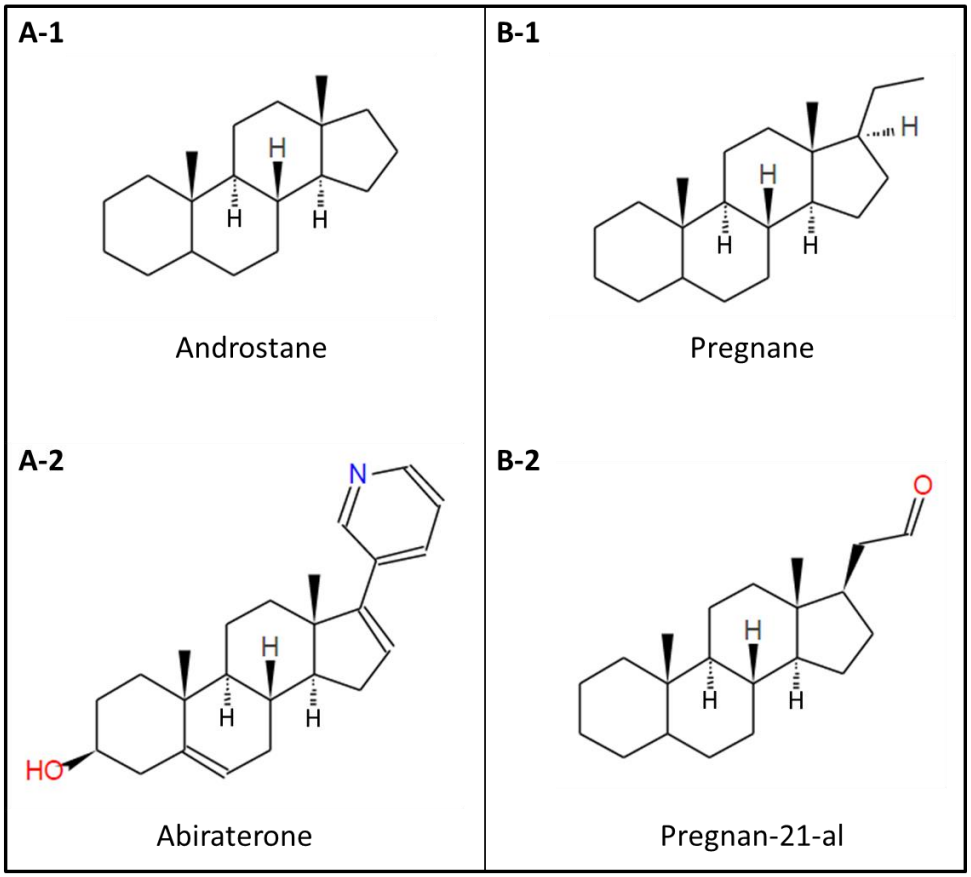


Figure 2.6: Chemical structures of androstane and pregnane compound classifications and a derivative of each. The structures of androstane (A-1) and one of its derivatives, abiraterone (A-2), and pregane (B-1) and one of its derivatives, pregnan-21-al (B-2) are shown. Structures were drawn using the JSME molecular editor.

2.3.1.3 Sources

The database currently contains compounds isolated from 145 different source organisms. Although it is estimated that 3000 species of plants are used for medicinal purposes in South Africa [47] and there are accounts of plant extracts being tested for various forms of activity [53–55], it is difficult to tell how many South African species (both plant and marine) have had compounds isolated from them. As such, it is difficult to assess the completeness of the

database from this perspective. Something that can be done to further develop this section of the database would be to link the source organism data to external data sources. A good example of this is PlantZAfrica [138], which was developed by the South African National Biodiversity Institute (SANBI). The website provides information about plants from across Southern Africa and so would be useful if the database is expanded to include compounds from this entire region. This would be helpful to further study these organisms and gain more information regarding where they occur.

2.3.1.4 Uses

Quite a large portion of compounds in the database have confirmed biological activity (Figure 2.7). Of those assigned, 40% displayed some form of anticancer activity. These were often broken down into a specific form of activity, such as cytostatic activity against HL-60 leukemia cells [139]. Compounds which displayed general cytotoxicity (i.e. against vero cells) were excluded from this grouping. Much like with the classifications, the uses can be divided into a hierarchy, and Figure 2.7 shows the compound uses at the highest level of this hierarchy. It is worth noting that when going through publications, only compounds that the authors considered to display significant activity were assigned use entries. Those that displayed low or no activity were not recorded. It may be worth recording that these compounds displayed no activity, so as to distinguish them from compounds that have yet to be tested for specific activity. Also, details about the way these activities were tested and actual values for each would also be helpful, though arguably ChEMBL should already have this information. The database would need to be modified in order to handle all this information, but it would enrich the data already present. An additional section that is currently being added is the known uses of the source organisms. Although some of the traditional uses may be anecdotal, a number of South African plant extracts have been shown to contain certain biological activities (refer to Section 1.4). Currently, the database houses

160 records of these uses, extracted from publications that isolate compounds from these sources. These have yet to be traced but to their original references, which is the next step that needs to be addressed for these entries.

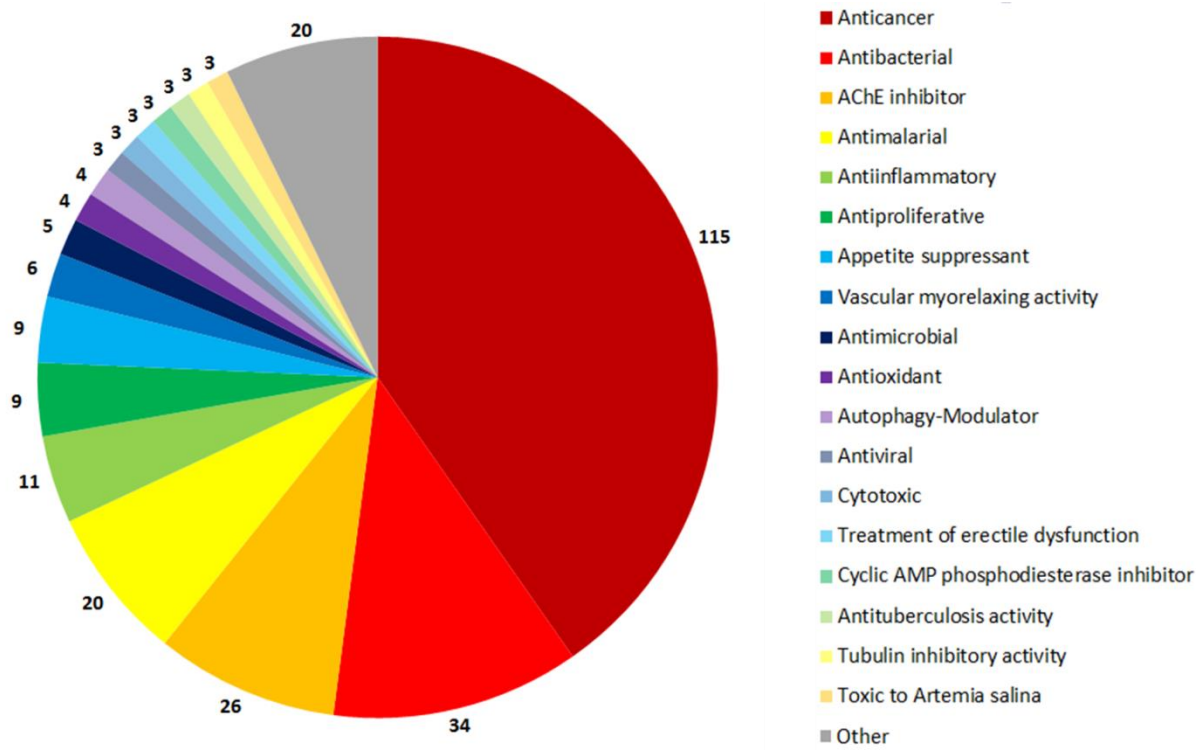


Figure 2.7: Distribution of compound uses within the database. The pie chart indicates the different uses recorded in the database, as well as the number of compounds with confirmed activity matching that use entry.

2.3.1.5 Other names

While searching for a compound in literature and other online databases, such as PubChem [72], it quickly becomes clear that an individual compound can be known by a number of different names. This can be challenging when finding the same compound in different publications. One of the goals when creating the database was keeping the information fully referenced. So if there was information from literature that was specific to that publication, the link between the compound and that information was also linked to the literature (Figure 2.3). This helps when addressing the different names of compounds in different publications, which can cause confusion. A simple example of this would be the compound isolated by

Sidwell & Tamm [140] and given the name autumnalin. This was soon changed to eucomnalin [140] because the name autumnalin had already been assigned to a different compound. In a review by Koorbanally *et al*. [141], the same compound is still referred to as autumnalin, as well as (E)-5,7-dihydroxy-6-methoxy-3-(4'-hydroxybenzylidene)-4-chromanone. The database keeps track of these changes and makes it easier to find a compound with a specific name or used to have a specific name.

2.3.1.6 Physicochemical properties

A common way to assess the drug-likeness of many natural products databases [67, 94, 95] is by using Lipinski's rule of five [121]. This involves assessing compounds, using four different properties: 1) the number of HBDs, 2) HBAs, 2) the cLogP value and molecular mass. The rule states compounds should ideally have no more than five HBDs, no more than 10 HBAs, a cLogP value of less than 5.0 and a molecular mass not exceeding 500 Da. These properties were calculated for compounds in the database. The number of Lipinski violations within the database is summarised in Table 2.1, each displayed in more detail in Figure 2.8.

Table 2.1: Lipinski violations of compounds within the database. Compounds are grouped based on the number of violations they make to Lipinski's rule of five. The number of compounds in each grouping is indicated, as well as what percentage of the total number of compounds in the database these account for.

Number of Violations	Count	%
0	373	62
1	112	19
2	84	14
3	30	5
4	0	0

Approximately 62% of compounds in the database were shown to violate none of Lipinski's rules and over 80% of compounds violate at most one of the rules. These are promising results, even when comparing to other databases. The properties used to define these violations are associated with the solubility and permeability of known drugs [121], which is why they give an indication of the "drug-likeness" of small molecules. The number of molecules with one or fewer violations greatly supports the potential of these compounds to be used for *in silico* drug screening studies. In addition to Lipinski's rule of five, other properties have been used to assess compounds in other databases, including molecular volume, TPSA and number of rotatable bonds (NRB) [67, 94, 95]. Enriching the current compound database with this information will also be of value.

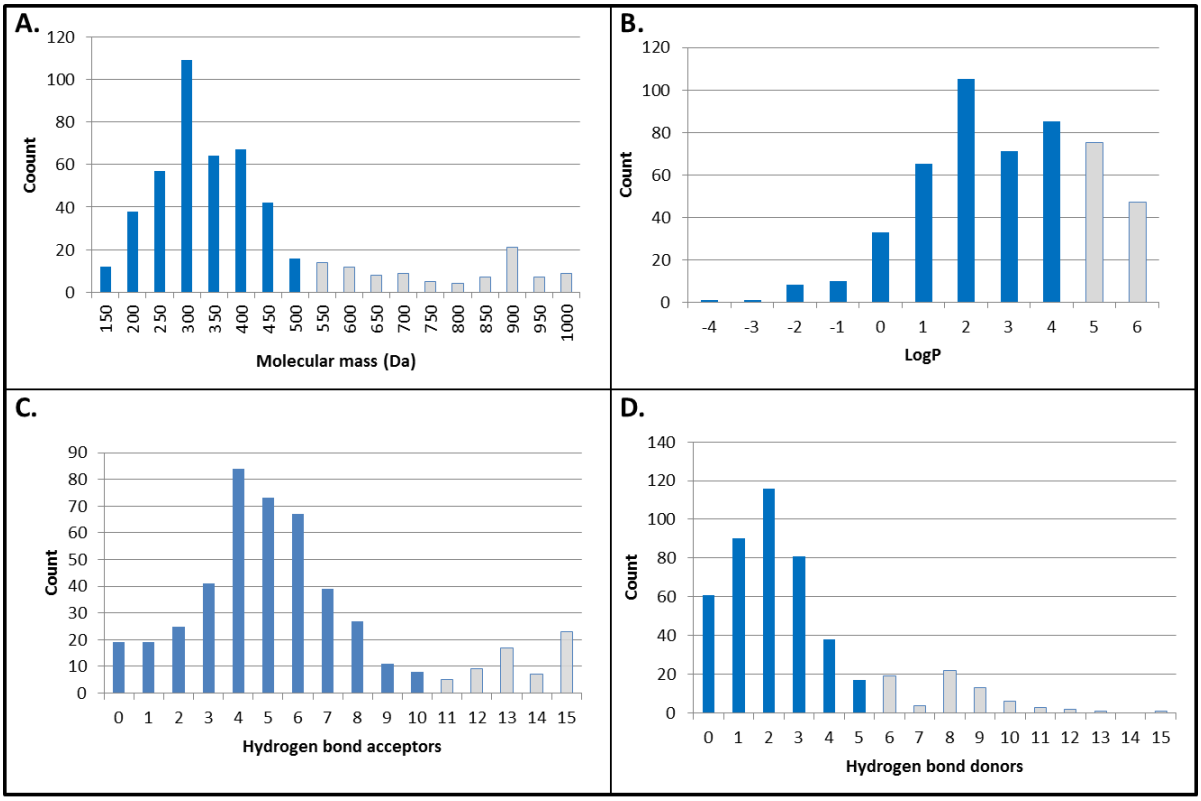


Figure 2.8: Physicochemical properties of compounds in the database, based on Lipinski's "rule of five". Bar graphs are show, indicating the number of compounds with specific values for each of the properties calculated when determining their "drug-likeness". These include A) Number of hydrogen bond acceptors; B) Number of hydrogen bond donors; C) LogP values and D) Molecular mass. Bars coloured in light grey indicate compounds with values above the cut-off, considered as Lipinski violations. Adapted from Hatherley *et al.* (in press) [70].

2.3.2 Error fixing

Apart from myself, three depositors assisted with uploading information to the database. These were David Penklar, Ngonidzashe Faya and Thommas Musyoka, each depositing a minimum of 100 compounds. As this was done manually, the process required steps to check for errors made while uploading. Firstly, within the construction of the database, a number of provisions were made to avoid redundancy. Fields such as DOI and CAS numbers required validators to be manually created in the **models.py** file to ensure these were unique if present. Scripts were written to find common errors that occurred when depositors uploaded compound information. These errors were identified by manually inspecting the entries of each depositor and are continually updated. Currently, the scripts checks for information that may have been omitted. For example, it checks to make sure that information linked to at least one reference, makes sure that references have at least one first author entry. The script also looks for formatting errors, such as unusual characters that should not be present in a SMILES string, as well as the format of external compound IDs and DOI numbers. These need to be entered in a specific format to ensure they can be used to link to the entry on external sites. Additionally, a link tester page was also created for entries with external compound IDs. After these IDs are entered by a depositor, the page automatically created links to the entry on the respective sites. These could be used by the depositor to quickly check that the ID was entered correctly and referred to the correct compound entry on the external site.

The SMILES input process has been revised to minimise potential errors. When a compound image is uploaded to OSRA, the structure is automatically drawn within a JME applet provided by the server. This is the precursor to the JSME applet, written in Java, rather than JavaScript [105] and could also be used to correct the structure, if inaccurate, as well as produce SMILES for the resulting compound drawn. The SMILES was originally checked by

using the Daylight Depict site (URL), which renders a 2D image of a compound if the SMILES is provided. The image produced by the SMILES was compared to the original compound image uploaded to OSRA by visual inspection to check the accuracy and make sure the SMILES string was valid. It also provided a quick means to make sure all rings were closed when the structures were drawn and that no hidden atoms were present in the structure, as could sometimes happen. There are, however potential problems with just using Depict to check for accuracy, especially with larger and more complex compound structures. Depict would not always return an image of the compound in the same orientation as the original compound image, making it difficult to compare the two structures by visual inspection. Also, it was easy to miss small changes between two structures this way, such as an additional heavy atom in long, aliphatic chains, or a substitution of a Carbon atom with Hydrogen, for example. Also, there was no quick way to tell if all known stereocentres were defined, without explicitly checking and comparing each. To remedy this, an additional step was added to the upload process. The Open Babel suite provides a program called obprop, which calculates certain properties of compounds, including the chemical formula and InChI. A quick way to tell if the compound was correctly drawn was to use obprop to calculate the formula produced by the SMILES string of the drawn compound. If the structure looked similar to the original 2D image and the formulas matched, it was considered safe to assume that the skeleton of the structure was at least correct. The InChI was then consulted to determine if the stereochemistry of the compound was fully specified. If not, question marks appear within the certain portions of the InChI string. This allows a depositor to quickly see if any chiral centres may have been missed when drawing the compound. Finally, if the PubChem ID of the compound was known, the InChI of the compound could be compared to the corresponding entry in PubChem. In addition to improvements in the methods used to

upload compounds, the data is timeously inspected by spot checks to see if errors may have been made or if information was missed when depositors were going through their references.

The generation of 3D coordinates for compounds also needed to be revised. This was originally done using Open Babel. However, there are parts of the Open Babel's 3D generation capabilities that are still under development [116]. When this was initially used, structures were checked by Dr Lobb for potential inaccuracy and corrections were made manually by him using Discovery Studio. This was also part of the original means of checking the SMILES input by depositors when capturing compound information. Once obprop was incorporated into the error-checking process, the InChI of the 3D structures produced could be compared to that from the original SMILES string used as input to generate these structures. Using this, it was found that just under 200 of the 3D conversions performed by Open Babel contained at least one stereocentre that was incorrect. Further investigation revealed that this was due to a specific type of substructure that the software struggled to handle. Although there is no official documentation that describes this problem, it is discussed online [142]. The way Open Babel generates 3D coordinates for compounds is a rule-based approach, which is used assign the backbone structure of the compound before the stereochemistry is fixed. The problematic substructure encountered was one where two rings join and the chiral atoms are bonded to three different ring atoms (Figure 2.9). These can not be fixed without breaking and reforming the bonds, which Open Babel is unable to do and leaves the substructure as is. To remedy this, the commercial software CORINA [118, 119] was used. Conversions made using this software appeared to retain the correct stereochemistry specified in the SMILES string. CORINA is used by the ZINC database, which indicates it should produce structures of sufficient quality for public use. The well-established online databases do used commercial software to ensure the quality of compounds within their databases.

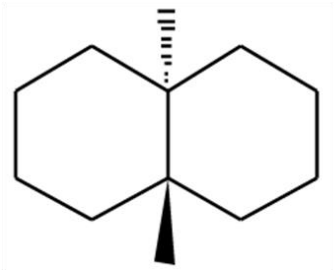
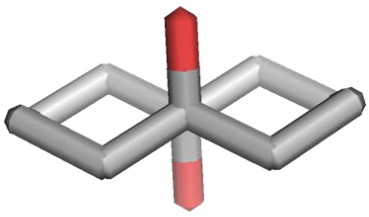
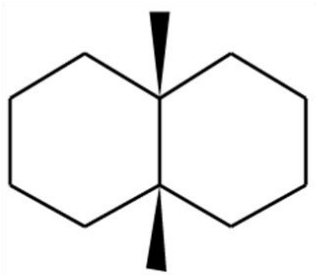
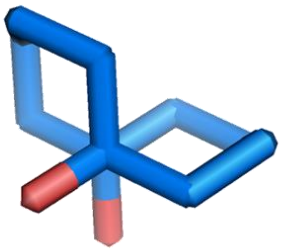
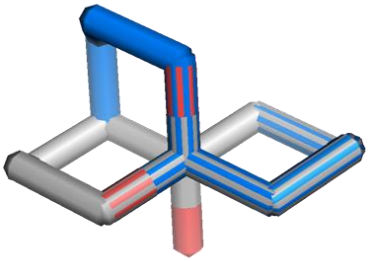
	2D	3D
trans		
cis		
overlap	**N/A**	

Figure 2.9: Fused ring stereocentres. The 2D and 3D of fused rings are indicated for both *trans* (grey) and *cis* (blue) structures, based on the bridgehead carbons. The bonds facing out of the ring are coloured in red.

2.3.3 The SMILES parser

The SMILES parser consists of a number of scripts that are largely still under development. Its current incorporation into the work presented is very limited, but with a great deal of potential to grow. At present it has simply been used to identify and count hydrogen bond donors of compounds in the database as part of the chemical properties calculations (Section 2.3.1.6). This involves looking at the number of Nitrogen or Oxygen atoms in a compound with free hydrogens. Since SMILES leaves hydrogens to be specified implicitly, these need

to be determined by looking at the number other heavy chain atoms attached to these Oxygens and Nitrogens, including those within rings, whilst considering bond types and the charges of these Oxygen/Nitrogen atoms, if present. These are fairly light-weight calculations that can be performed for all 600 compounds within the database virtually instantaneously, which make them useful to be incorporated into a simple Python script. In future, the SMILES parser will be improved and further developed to increase its functionality. The most immediate goals involve use in compound image generation and enhancing substructure searching.

2.4 Conclusion

Presented in this chapter are the design, construction and population of a compound database that focuses solely on natural products isolated from South African organisms. The retrieval of data was, as far as possible, linked to external references, which can be looked at in future to both verify the information extracted, as well as go back to the original work to review details not captured for the purposes of the database presented. Apart from the chemical information, data concerning other names, source organisms, uses and most recently anecdotal uses of source organisms were captured, each linked to at least one external reference where possible. The current content of the database consists of 600 compounds, from 170 different references pertaining to 143 different source organisms. There are also 284 different records of compounds showing promising activity and 160 records of anecdotal uses for the source organisms, as used in traditional medicine. There are currently several plans to enhance the current data contained within the database concerning source data, use data, physicochemical properties data. In addition to this, the number of compounds in the database will continue to grow, further expanded by including simple derivatives reported in literature, as well as potentially increasing the borders of the information content to include

all countries within Southern Africa. Data quality in the database is incredibly important and this will continue to be monitored and means to improve error checking capabilities developed to ensure high quality data. Calculations performed on the compounds extracted indicate that over 80% have promising physicochemical properties, based on Lipinski's rule of five, for future inspection in drug development. To assist with this, 3D structures of all compounds have been calculated and minimised for future *in silico* screening studies. To take full advantage of this, the next step, presented in Chapter 3, involves developing a means of sharing this data with the research community. This ensures that its use is maximised and the compound information isolated can be used for screening against a variety of potential drug targets, not just those of interest to our own research group.

Chapter 3 The SANCDB website

In the previous chapter, a database of compounds isolated from organisms found in South Africa was established. In order to make this information easily accessible to the general research community, a website was created as an online interface to the database. The website was named the South African Natural Compounds DataBase, or SANCDB. It allows compounds to be searched for based on all information gathered on each and contains links to data from external sites. Apart from a graphical interface, a web API was also established, so that information can be retrieved without having to visit the SANCDB web page. Also included is a compound submission pipeline which will enable other researchers to openly contribute to the database. This will help develop SANCDB into a community driven front end to information gathered concerning natural products research in South Africa. The work presented in this chapter and Chapter 2 was recently accepted for publication [70]. This website will be made available at https://sancdb.rubi.ru.ac.za/ as soon as it is published.

3.1 Introduction

While it is not uncommon for researchers to have their own libraries of chemical compounds [67, 143–145], this information can often be kept within a research group or otherwise be difficult to access. In bioinformatics, open source databases such as the PDB and GenBank were found to be incredibly beneficial in progressing both molecular and structural biology [146]. Presently, there are a number of online chemical compound databases and the majority of data lies with commercial databases. For example, the commercial SciFinder houses 93 million compounds, compared to PubChem, the largest public chemical resource, which contains 53 million [147]. Recent trends have indicated, however, that this may not be the case indefinitely [147].

3.1.1 Commercial compound databases

Commercial databases pride themselves on having high-quality data, as well as a great deal of information not found in freely accessible databases [146]. SciFinder [103] is probably the most well-established and revered commercial compound database and often either SciFinder or CAS are referred to when discussing commercial databases [146–148]. Reaxys [149] and NAPRALERT [90] are other good examples which are used in present-day research [94, 150–152]. Although Reaxys has no official publications, in literature it has been used to retrieve over 118 000 unique NP chemical entities, not found in other databases [94]; for retrieving physicochemical properties for a set of compounds [150] and searching for literature and patents regarding the chemical constituents and bioactivity data of a specific species of plant [151]; just to name a few of its uses. In 2008, SciFinder was making over $250 million annually, charging for their services and are able to use this to hire skilled chemists to ensure the high quality of their data [146].

3.1.2 Free-access compound databases

ChemSpider [153] is a good example of an open source collaboration. Developed by the Royal Society of Chemistry (RSC), the database contains over 30 million unique chemical structures from over 500 different online resources and even grants access to information from commercial databases [154]. Freely accessible databases often require a great deal of external funding, such as PubChem (NCBI), which is funded by the NIH [74], and the RSC is a group with over 50 000 members [154]. Both groups are funded for the development of scientific knowledge rather than commercial gain. An additional set of chemical data includes those from open source initiatives. These receive no funding but are managed by the chemistry community, as with Blue Obelisk organisation [155].

3.1.3 Web frameworks

The internet forms a crucial and increasingly common means of gaining access to scientific information, and this is no exception to information that is pertinent to chemists [146]. The means of developing websites has gone through a number of stages to make handling increasing large quantities of data more practical and secure [156]. Web pages can be considered as files written in Hyper Text Markup Language (HTML), styled with Cascade Style Sheets (CSS), which together describe the content of a web page [157]. HTML is the language web pages are written in and is interpreted by a web browser [157], while CSS is used to specify how different parts of a web page look and is often contained in a separate document to an HTML file [158]. Additionally JavaScript, which allows the HTML and CSS within a page to be manipulated, as well as Ajax, which handles data requests without the page needing to reload every time, have become important to web development [159]. The most common way to create dynamic data-driven websites is to combine MySQL with PHP, a scripting language that can be embedded within web pages [156, 159]. PHP is easy to learn and use, which contributes to its popularity, but can just as easily introduce security risks into websites [156]. The latest generation web frameworks, such as Django and Ruby on Rails, were designed such that HTML content could be generated dynamically, but also provide features which automatically guard against mistakes that compromise the security of a website [156].

3.1.4 Model-View-Controller design paradigm

Model-View-Controller (MVC) is an architectural design implemented when developing an application [160]. It allows the application to be separated into three independent parts, which can each be altered without affecting the others [156]. These are named the model, view and controller. The model contains the so-called “business logic” of an application. Here the structure of the data is designed, as is the means by which it may be accessed or altered. The

view is simply the part of the application that the user sees and interacts with, otherwise known as the user interface (UI). Finally, the controller connects the model to the view, allowing communication between the two [160]. The relationship between these three is such that the user makes a request via the view. The controller sends this request and any accompanying information to the model, which handles the request. The response is sent back by the controller and used to update the view.

3.1.5 Model-Template-View

The Django web framework has a slightly different interpretation of the MVC paradigm. The framework itself acts as the controller, since communication between the model and the view is handled automatically. The model layer only defines the data and how it may be manipulated. There is a template layer describes how the data should be displayed and a view, which houses the logic of the application and handles communication between the models and the templates. This is considered to be a Model-Template-View (MTV) design, but otherwise operates on the same principles as the MVC design paradigm [156].

3.1.6 Web servers

A web framework provides a means of designing a web application , but does not enable information to be sent and received over the internet [156]. This is done by a web server [161]. When a user accesses a web page, it will have a specific Uniform Resource Locator (URL), e.g. https://www.google.co.za. A web server is run on a host machine and uses the URL to decide what information is sent to the web browser and presented to a user. Apache is by far the most widely used webserver. It is designed to be secure, to handle multiple requests, and to support a number of different file formats when information is sent over the web [161].

3.1.7 Proposed work

The previous chapter describes the establishment of a database of South African NP database. Presented in this chapter is the use of this database to create an online interface which has been named, “SANCDB”, through which this information may be accessed, filtered and downloaded. The layout of the website is described along with the process that by which users may search for information, download content or follow links to access additional data.

3.2 Methodology

3.2.1 Website name and logo

In order to allow users to easily identify the site, a name and logo were created. The name "SANCDB" was chosen, which stands for the **S**outh **A**frican **N**atural **C**ompounds **D**ata**b**ase. The name was chosen with the criteria that it was descriptive of the contents of the database and also that it had not been used elsewhere. This second criterion was checked by means of a Google search [113]. Secondly, a logo was designed and created using simple shapes provided by Microsoft PowerPoint.

3.2.2 Layout of the website

The SANCDB website was created using a CSS bootstrap template [162], which had previously been modified by David Brown from the Research Unit in Bioinformatics (RUBi), Rhodes University; and further modified for use in this project (Appendix D). Django was used to specify the content of each page and handle calls to the database. JavaScript and Ajax were used to provide functionality on each page, after these had been rendered by Django.

3.2.1 Web API

A RESTful API was also included using the Django REST framework. A set of Serializer classes were created, based on the Django models describing the database. These serializers simply describe the layout of the data to be transferred, much like the classes in **models.py** do. API View functions were written for a variety of URLs to account for different possible usage scenarios. The API Views, URLs and Serializer classes are shown in Appendix A; Files: **serializers.py, urls.py** and **views.py**.

3.2.2 Website pages

3.2.2.1 Search pages

The search pages generally comprised four main components: 1) Initial logic from the view function, which renders each page; 2) Template scripts and forms, which present search options, capture search queries and make Ajax calls; 3) A second view function which queries the database for the information requested and returns the data retrieved to the template and 4) Scripts which updates the template and page with the search results. The Python library, pylibmc, was used to cache common database calls. The logic incorporated into these pages works as follows.

3.2.2.1.1 Rendering view function logic

For pages with browse sections, a database call is made which retrieves all search entries for that page and sorts them alphanumerically. These are stored in a list of tuples, with the name of the entry as the first item in the tuple and the ID of that entry as the second item in the tuple. The template for the page is then rendered with this information.

3.2.2.1.2 Template forms and search logic

Template scripting was done using JavaScript, using the jQuery and jQuery UI JavaScript libraries. The code used to perform these searches was specific to different search types (Appendix E; Files: **sanc.js**, **sanc_advanced.js, sanc_physico.js, sanc_refs.js and sanc_search.js**). Firstly, for pages with browse sections, the data from section 3.2.3.1.1 was used to create a list of names, relevant to the page and divided into sections, based on the letter they started with. A set of buttons was created, each of which with a single letter of the alphabet displays the list of names that start with that letter. The ID associated with each name was used to create a link to a page containing data specific to that entry.

The jQuery UI autocomplete function was used to create search bars which update as the user types, showing all possible search terms that contain the letters typed so far. The data used to populate these was retrieved using an Ajax call to cached data. When the search button is clicked, the search terms are converted to JSON and an Ajax call is made.

3.2.2.1.3 View search logic

The view function was written to convert the JSON from the Ajax call using the json Python library and queried against the database. Django's Q objects were used for queries involving multiple search criteria. The results are included in a list, converted to JSON and sent as a response to the Ajax call.

3.2.2.1.4 Template display of results

The results of the Ajax call are tabulated, with links to the compound summary pages and functionality to download specific compounds.

3.2.2.2 Compound summary page

Included in the URL for a given compound summary page is the ID of the compound, as it appears in the Compounds table of the database (refer to Section 2.2.1). This ID is used as a parameter in the view function to retrieve information for that specific compound. The SANC ID for the compound is prepared, as are the list of other names for the compound, the compound formula, the reference citations, as well as the links to PubChem and DrugBank (refer to Section 3.2.3). This information is used when rendering the template (Appendix F; File: **comp_page_2.html**).

3.2.2.3 API documentation page

The API documentation page was generated using the Django REST Swagger documentation tool. The information specified in the documentation was included in each API view function.

3.2.3 Data preparation functions

A number of functions were incorporated into the Views file for the SANCDB application. These simply prepared the data in a specific format so it could be used within certain pages. The following were all incorporated into the compound summary page.

3.2.3.1 SANC ID

The SANC ID of each compound is an identifier, used to name the compound within the database. Each SANC ID is the same length, consisting of the prefix "SANC", followed by a five-digit ID number, which reflects the compound's primary key ID in the Compounds table of the database. This was used when naming compounds, as well as files associated with the compound. For example, compound SANC00101 can be searched locally in the database using its ID '101' and coordinate files of the compound will be named SANC00101.pdb, SANC00101.sdf, etc.

3.2.3.2 Compound names

When the names of compounds were captured, unusual characters, such as 'α' and 'β' were converted to '[alpha]' and '[beta]', respectively when stored in the database. These were converted back to the original characters for display on the summary page. The script also includes steps to standardize the way a compound is written, with respect to capital letters, italics, etc., since these were provided in different ways by different references and also handled differently by depositors. The script identifies 'whole words' within the name, based on the presence of spaces, dashes, commas etc. These are compared to a pre-rendered list of terms, for example, "*trans*" and "*cis*". If the word is not in this list and it is the first word in the name, it is capitalised. If it is in the list, depending on the term, it will generally be italicised and not capitalised. This is done for all captured names of the compounds and the script is updated based on new names and exceptions encountered as compounds are uploaded to the database.

3.2.3.3 Citations

These are just the way a referenced work is displayed on the summary page. The format is fairly simple and follows the pattern of 1) First author, 2) Author suffix, 3) Year, 4) Reference title. The Author suffix, changes based on the number of authors listed in the reference. For a single author, there is no suffix. For two authors, the suffix "and [Author 2]" is used, and for more than two authors, "*et al*." is used.

3.2.3.4 External links

The links to external databases include the URLs of compound entries in other databases, including ZINC, PubChem, ChEMBL and DrugBank. The URL was constructed based on the ID of the entry in the external database, as these have a standard format in each external database. The links were automatically created, based on the values stored in the database.

3.3 Results and discussion

3.3.1 Website layout

The general layout of the site is displayed in Figure 3.1. The template divides the website into three portions: 1) The navigation section; 2) The sidebar and 3) The content section.

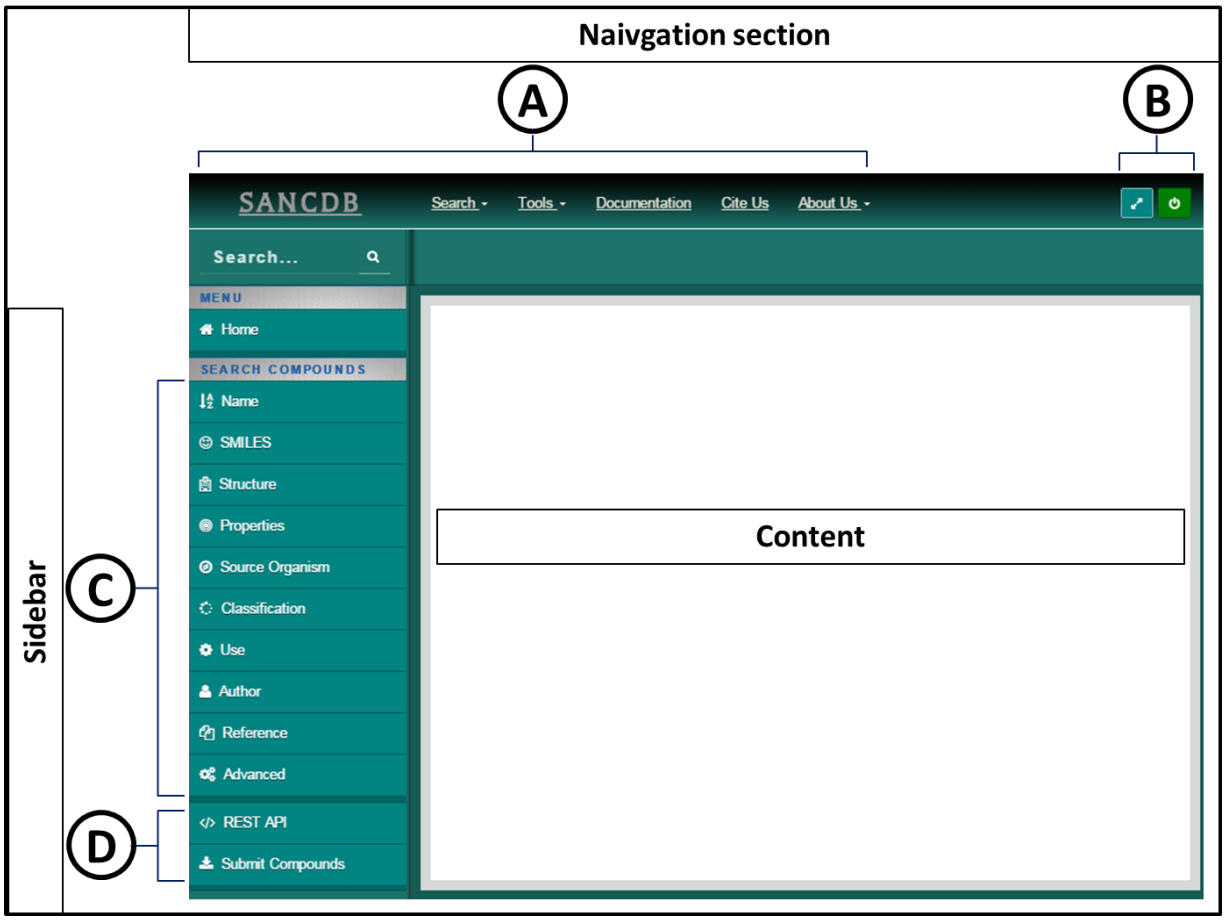


Figure 3.1: Layout of the SANCDB web interface. The general layout of pages within the SANCDB website is shown, with the different sections of the template used, labelled. The content section is the only part of the page that changes, displaying content that is relevant to the specific page accessed. The navigation section and sidebar can be sub-divided as follows. A) Navigation links; B) Login/logout button and widescreen button; C) Compound search links; and D) Additional links.

3.3.1.1 The navigation section

This contains additional links and drop down menus (Figure 3.1 A), as well as the login/logout button and a button which shows/hides the sidebar (Figure 3.1 B). The name of the site, SANCDB, is displayed in the left corner of the screen and links to the homepage.

The rest of the navigation links summarise the sidebar section into a series of dropdown menus. This section also contains links not found in the sidebar section of the screen.

3.3.1.2 User login

User login and signing up is handled by an application created by David Brown. A user needs to provide their username and password to log in and use the website. If the user does not have an account, they can sign up for free. The signup screen has been modified for use is the SANCDB website to capture the name and university/organisation of the user. As the website aims to be freely accessible, users will not be obligated to sign up or log in to browse the website content and download data. Login will be encouraged, however, as a means to simply help us keep track of who is using the website. This will also allow sessions to be geared towards specific users. For example, the site could keep track of usage history, so users could quickly retrieve compounds they downloaded previously or be notified if similar compounds are uploaded. User login will be required for compound submission (Section 3.3.5), as these entries may will need to be linked to the user's account.

3.3.1.3 The sidebar and page content

The sidebar contains quick links, commonly used in the in the website, including links to the different compound search options (Figure 3.1 C) provided, as well as tools associated with the site (Figure 3.1 D), each described in more detail below. The tools section is reserved for additional content and functionality that might be included in the site in future.

3.3.2 Searches

The SANCDB website currently provides ten different search options to allow users to find compounds of interest. These include searches based on 1) The name of a compound; 2) SMILES; 3) Structure; 4) Properties; 5) Source organisms from which a compound has been isolated; 6) Classification; 7) Uses; 8) Authors who have isolated or worked with a

compound; 9) Referenced works that pertains to a compound and 10) Combining multiple search criteria (Advanced search).

3.3.2.1 Text-based searching

Most pages involve searching simple terms to find compounds, based on a specific attributes. To describe these types of search pages, the Source Organisms search page will be used as an example (Figure 3.2).

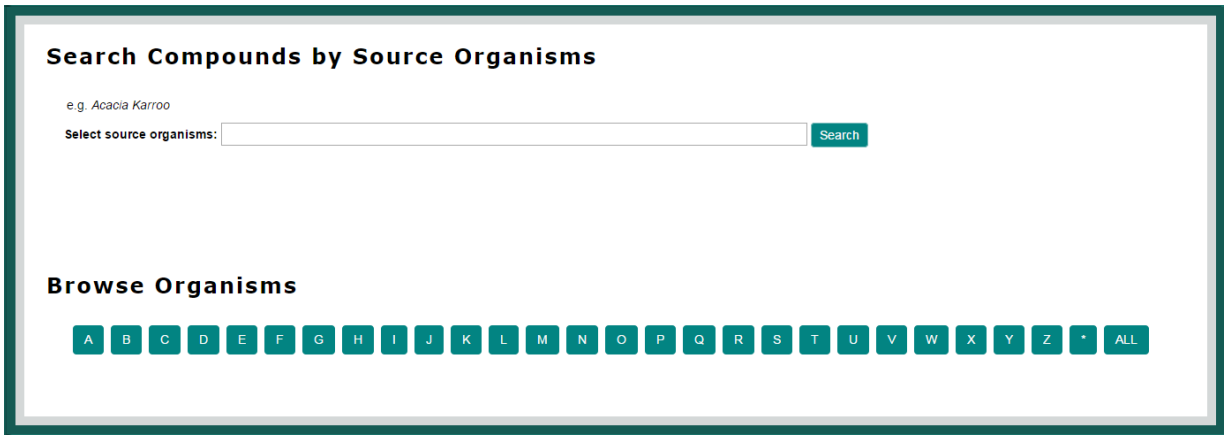


Figure 3.2: Source Organisms search - First view. The content section of the source organisms compound search page, as indicated in Figure 3.1, is shown. This page provides ways to search for compounds from specific organisms (top) and browse through the set of source organisms currently housed within the database (bottom).

These search pages are split into a Search section at the top and a Browse section just below this. The browse section of the page (Figure 3.3) contains buttons indicating the letters A-Z, as well as an asterisk ('*') and an 'ALL' option. As soon as one of these buttons is selected, the Search section falls away. A 'Show Search' button is presented in the top right corner of the page, which will show the Search section again. While browsing, the user may select any of the lettered buttons mentioned and all entries starting with that letter will be listed. The asterisk option is for entries that do not start with letters or begin with special characters, such as α or β. Clicking on any of the entries listed will direct a user to a page of all compounds associated with that entry.

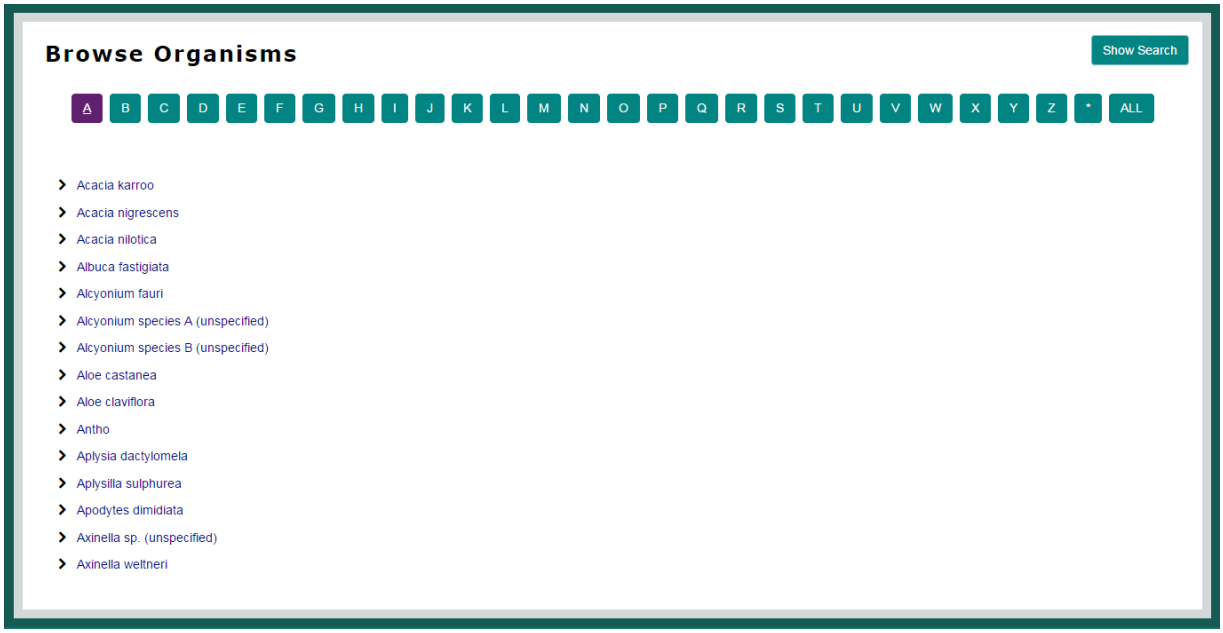


Figure 3.3: Source Organisms search - Browse section. This interface is shown when the browse section of the page is engaged. The search section is hidden, but can be returned by clicking in the 'Show Search' button. The source organisms are those that start with the letter 'A', displayed when the button for this letter is clicked on.

The Search section simply contains an input box in which users type in their search query. An autocomplete function has been employed, which indicates entries containing the letters input so far. Users may select search terms from the autocomplete list or use a semicolon (';') to separate different search terms. Multiple search terms can be input this way. Once again, when the 'search' option is clicked on, the Browse section falls away and is replaced by the 'Show Browse' button in the top right corner of the page. The actual search is handled by an Ajax request, which queries the database and displays the results without having to reload the page. The search results are tabulated as shown in Figure 3.4. The way the results are tabulated varies across the different pages, depending on what information is available. The SANC ID of each compound will be displayed, which links to the summary page for that compound (Section 3.2.2.2), along with its image and a check box to select it for download. If possible, the phrases matched when querying the database will also be displayed. For example, in Figure 3.4 the results are shown for a Source Organism search matching the

phrase “eucomis”. Compound SANC00221 ((R)-Scillascillin) has been isolated from both *Eucomis schiffi* and *Eucomis humilis*, so these are listed under ‘Organisms Matched’. The tabulated results have check boxes to select specific compounds for download, as well as a ‘select all’ option, which selects all check boxes for this purpose. The download process (Figure 3.5) allows compounds to be downloaded in Molfile (MOL2), PDB, SDfile (SDF) or SMILES format, as well as minimised compounds in PDB format. Once one of these options is chosen, all compounds selected are set for download as part of a .ZIP file.

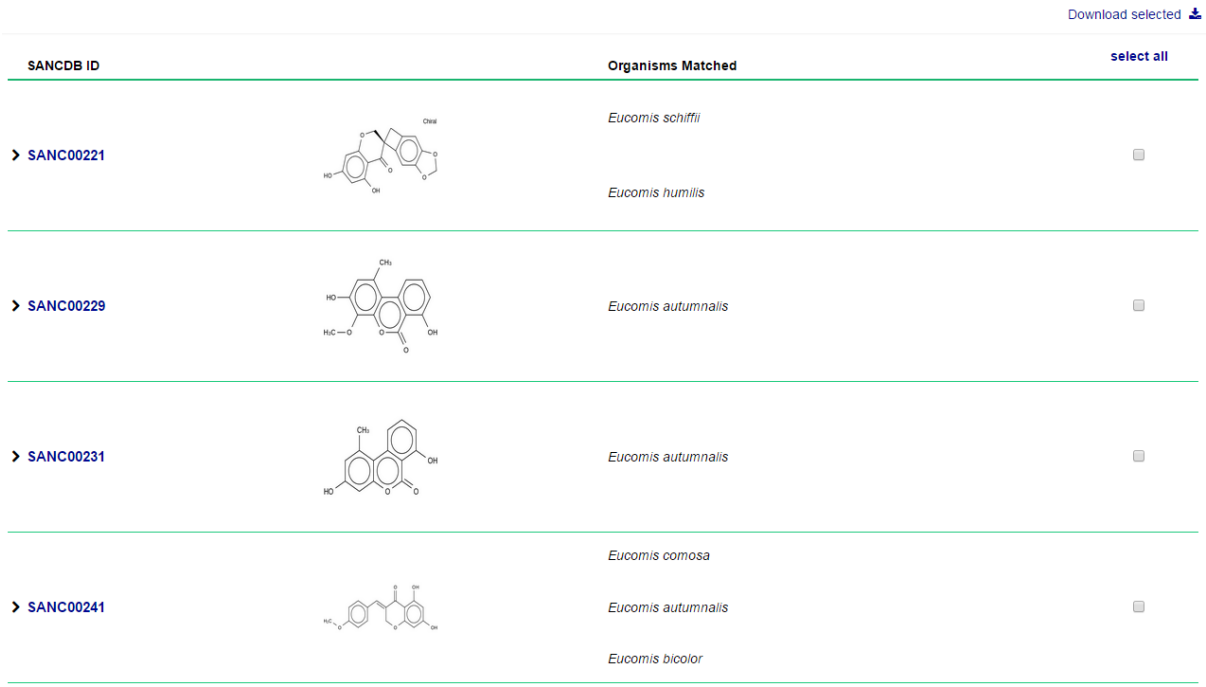


Figure 3.4: Source Organisms search – tabulated results. Search results are shown for searching using the query ‘eucomis’. Compounds are tabulated with their 2D images shown, as well as links to their summary pages. All sources that match the search criteria are also shown, and the check boxes on the right can be used to select compounds from this list for download.

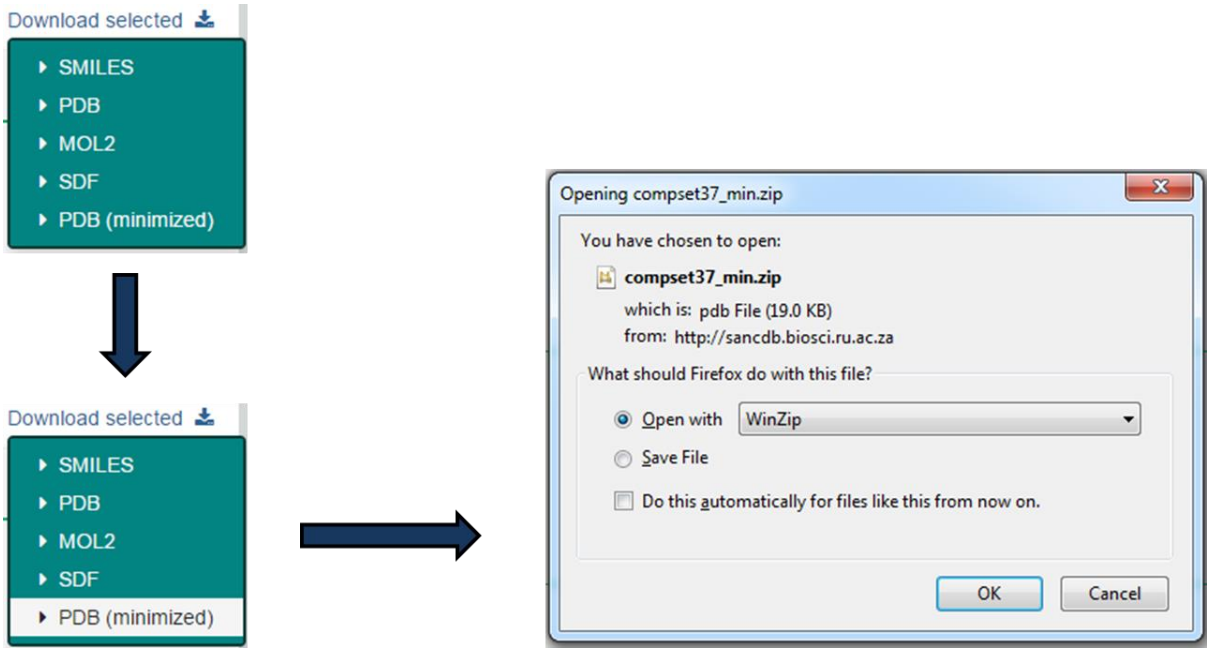


Figure 3.5: Downloading process. The download process is summarised. After compounds of interest are selected, the 'Download selected' button can be clicked to display a dropdown list of file formats in which the compound set can be downloaded. Once one of these is selected, the download process is initiated and compounds are downloaded as a .ZIP file.

3.3.2.2 SMILES- and structure-based searching

The SMILES search page and structure search page are similar in that they both use SMILES to query the database for matching compounds. The structure-based search is, however, more user-friendly as it employs the JSME applet (Figure 3.6). This allows users to simply draw in a chemical structure, substructure or functional group of a compound they might be interested in. The advantage of this is that the user requires no working knowledge of SMILES and can visualise the desired chemical entity before they search for it. This kind of search has become standard in chemical web databases. For example, ZINC [78] and NuBBE [95] both use the WebME drawing tool from Molinspiration [163], whereas ChEMBL offers a number of different drawing tools for this purpose [75]. These pages perform substructure searches by default, using Open Babel. In future, options will be provided to search based on similarity, again using Open Babel.

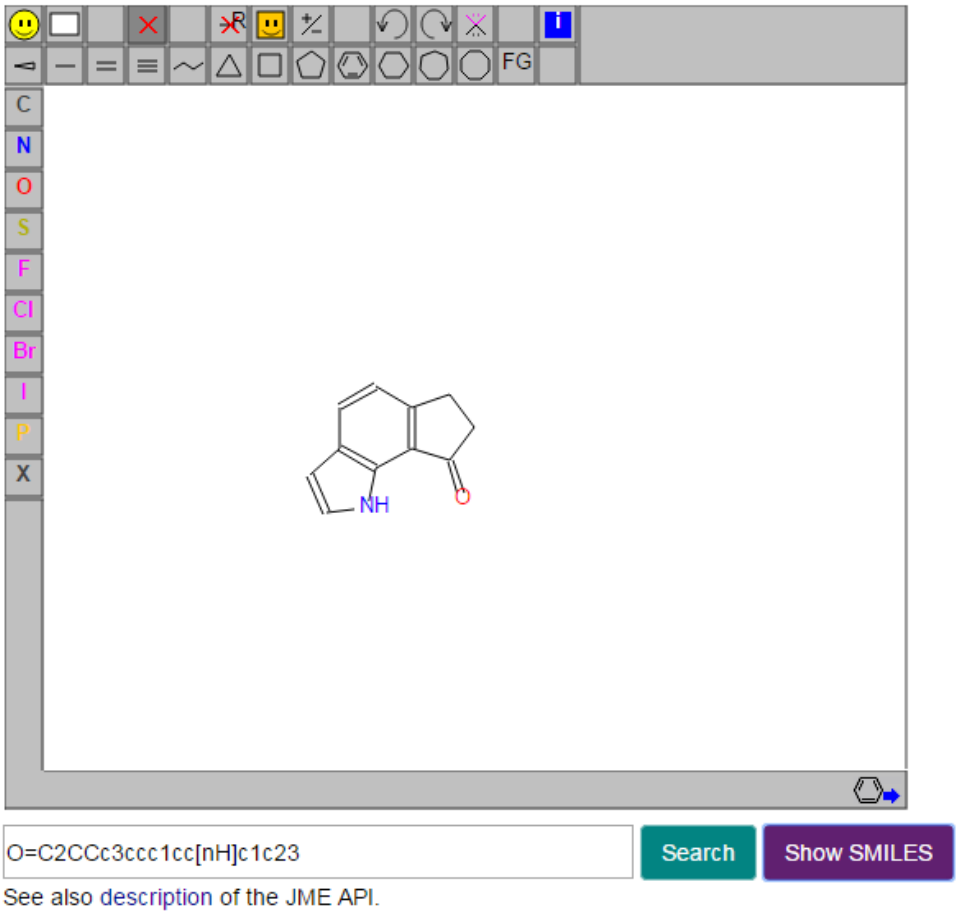


Figure 3.6: JSME applet. The structure drawing applet is indicated with a substructure drawn in. By clicking on the 'Show SMILES' button, the SMILES for the compound gets displayed, as shown. By clicking the 'search' button, the database will be queried for all compounds that contain the substructure drawn.

3.3.2.3 Properties search

The Properties search page allows users to search for compounds based on the physicochemical properties that have been calculated for these compounds. Currently, these are based on Lipinski's rule of five [121] include molecular mass, LogP, number of hydrogen bond acceptors and donors (Section 2.2.7). Slider bars are provided on this page in order to allow users so search for compounds with properties that fall within a specific range (Figure 3.7). This section will be expanded as the means to calculate additional properties become available.

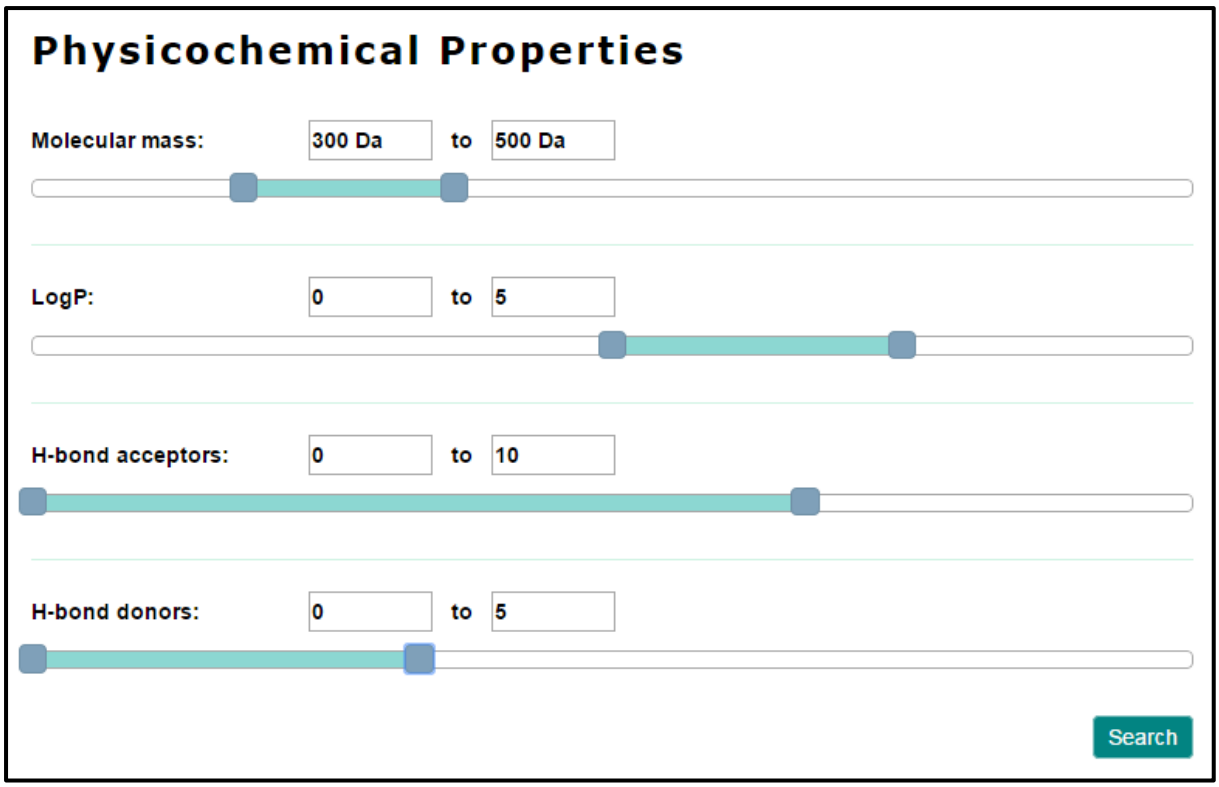


Figure 3.7: Properties slider bars. The search slider bars are shown for the Properties search page. The sliders can be used to set a specific range of values for mass of the compounds, cLogP, HBAs and HBDs. When the 'Search' button is clicked, the database will be queried to find all compounds with properties that fall within the ranges specified.

3.3.2.4 Reference search

The Reference search section is a more advanced version of the simple text-based search pages. As with these, the user can browse different references by their title. However, as with compounds, references have a number of different attributes by which they can be searched against. As such, the Reference page allows users to search for references based on Title, Year, Authors, Journal, Source Organisms and Uses. These can be distinguished using radio buttons, to allow users to decide which attribute they would like to search by. As with other text-based search pages, an autocomplete function is incorporated to allow users to filter their search options as they type. The list of options provided by this function changes based on which search option is selected. This is demonstrated in Figure 3.8. The result of a search is still a tabulated list of compounds. The title of the reference matched for each compound is

displayed, as well as the search phrase that was matched when querying the database (i.e. the name of the author or journal etc. searched).

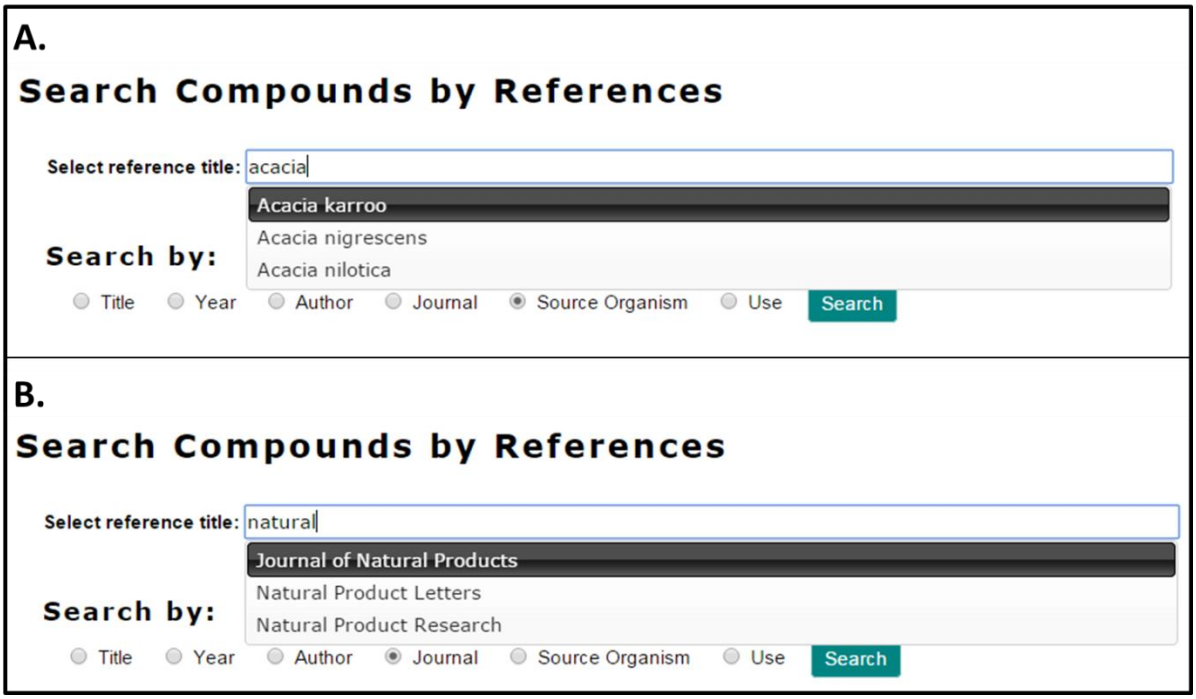


Figure 3.8: References search bar. The References search page is shown. References can be searched for based on Title, Year, Author, Journal, Source Organism or Use. These are selected using the radio buttons shown. When a radio button is selected, the content of the autocomplete function id updated to be relevant to the search option. For example, in (A) a set of source organisms make up the search list and in (B), the names of journals are used.

3.3.2.5 Advanced search

The Advanced Search page combines text searches across different compound attributes that could be searched individually through the various other text-based search pages. Currently, compounds can be searched by combining search queries for source organisms, compound classifications, uses, year of publication and author name of publication. These are presented as five separate search bars (Figure 3.9), each with their own list for the autocomplete function. Entries from different search bars are treated as 'AND' queries and those from the same search box are treated as 'OR' queries.

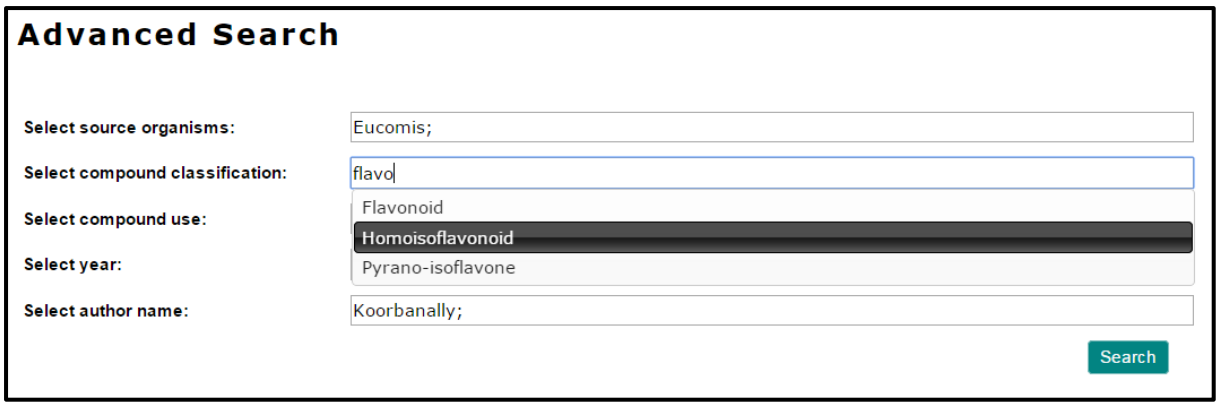


Figure 3.9: Advanced Search bar. The Advanced search page is shown, which allows compounds to be searched for using a number of different criteria. Each search bar has its own autocomplete function.

3.3.3 Compound summary page

The compound summary page contains all information in the database about a given compound. The header of the page is the SANCID of the compound, followed by basic information, such as the compound's name and formula. If available, external IDs are given for entries of the same compound in other online databases, along with links to the compound entry pages in those databases. Two different graphic representations are provided for each compound (Figure 3.10). By default a 2D image of the compound is shown, as produced by the Indigo toolkit. A '3D View' button is provided which shifts this to a 3D representation of the compound without hydrogens. The compound shown is a PDB file rendered in GLMol. This uses WebGL and allows the compound to be freely rotated on screen to get a better look at it. A '2D Image' button is also provided to switch back to the 2D compound image representation. Just below the compound image is a download section. This works in the same way as the tabulated download system, except a single file will be downloaded for the compound in the specified file format, rather than a compressed .ZIP file. Below this is the rest of the information available for the compound, including its SMILES, references and recorded classifications, other names, source organisms and uses. The references section also contains links to the online resource containing the reference. This is done based on the

references DOI number and usually goes to the journal's website entry for that specific article.

3.3.3.1 External links

As each compound was added to SANCDB, the external databases ChEMBL, DrugBank, PubChem and ZINC were searched to see if they contained the compound. PubChem searches yielded 141 compounds, which was the greatest number found in these databases, followed by 96 compounds from ChEMBL. This is understandable, since PubChem is the largest publically-accessible compound database [147]. Only 19 compounds were found in ZINC and nine from DrugBank.

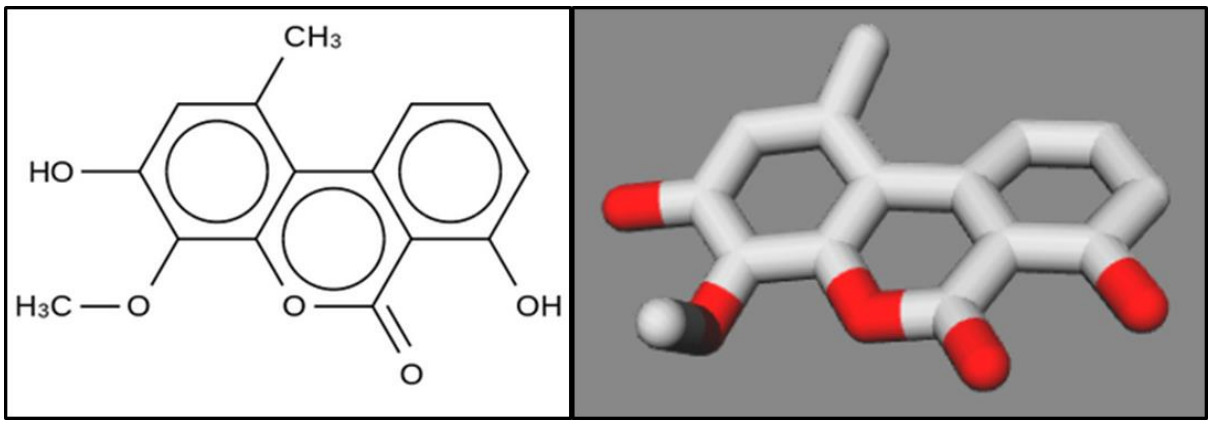


Figure 3.10: Compound 2D image and 3D representation. The 2D image and 3D representation of compound SANC00229 (Autumnariniol) is shown. The 2D image was created using the Indigo toolkit and the 3D coordinate file for the compound was rendered using GLMol. Both of these depictions can be viewed on the compound summary page.

3.3.4 Web API

The web API allows content of the website to be accessed and retrieved without needing to use the web interface. This is also easier to incorporate into external applications that do not have access to the database used by the SANCDB website. The web interface of SANCDB provides a link to the documentation for the web API, generated using the Django REST Swagger tool. There are two different ways to retrieve information from the web API. Firstly, there are search URLs (Figure 3.11), with the "/api/search/" prefix which provide similar

search options to the web interface. This accepts search terms to query the database and find, for example, all records of compounds from the source organism *Acacia karroo* (URL: /api/search/compounds/source/acacia%20karroo/). These search URLs were designed to return information about each compound, as well as references associated with them. The information is returned in JSON format, structured as it appears in the database.

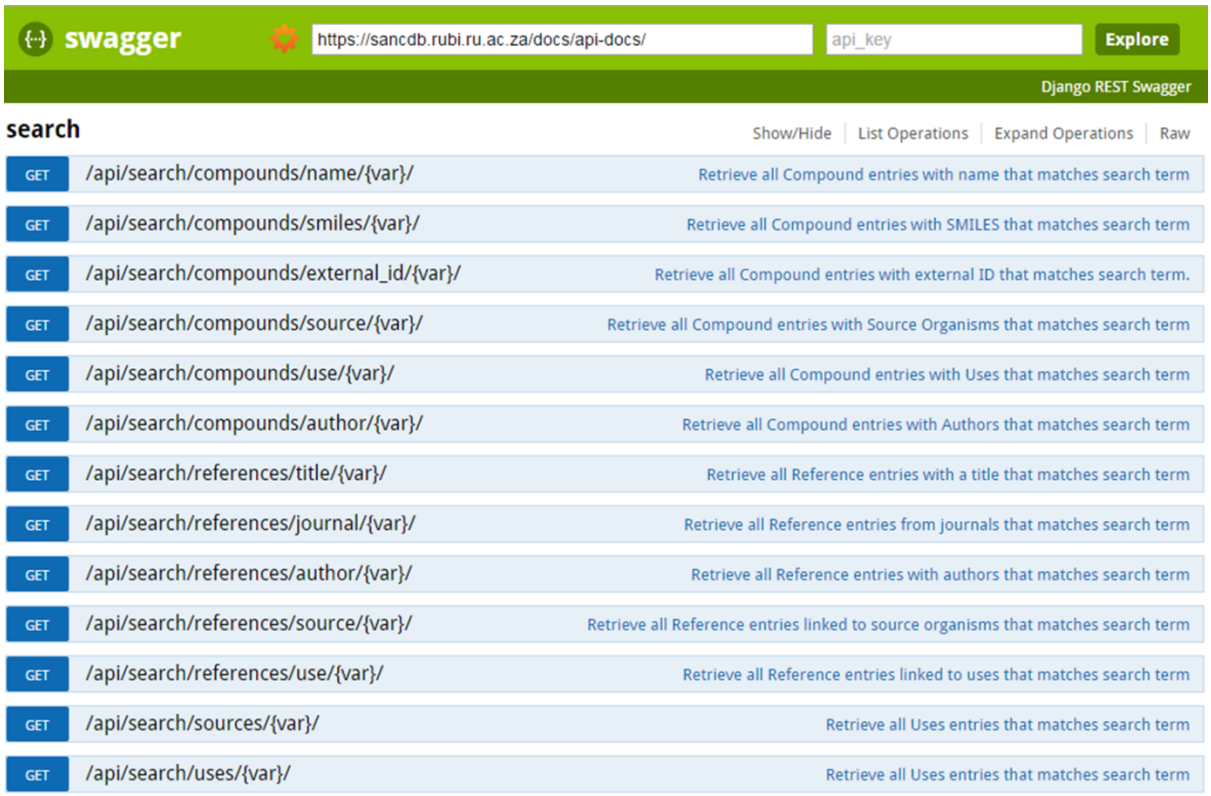


Figure 3.11: Django REST Swagger documentation page, showing search options for web API.

The second way to retrieve information is to use the specific retrieval URLs (Figure 3.12). These are useful if the IDs are known for specific entries, which are included in the data returned when using the search URLs. These URLs currently apply to authors, classifications, compounds, other names, references, sources and uses. To demonstrate how this may be used, the source organism, *Cephalodiscus gilchristi*, can be considered. This organism has the ID "100" in the Sources table of the database. To retrieve information about the structural classifications of compounds found in this organism, the URL,

/api/sources/100/classifications/, could be used. To find all references in the database associated with this organism, the URL, /api/sources/100/references/, would be used. As such, by combining the search URLs with the retrieval URLs, a great deal of specific information can be retrieved for different entries in the database, without having to use the SANCDB web interface.

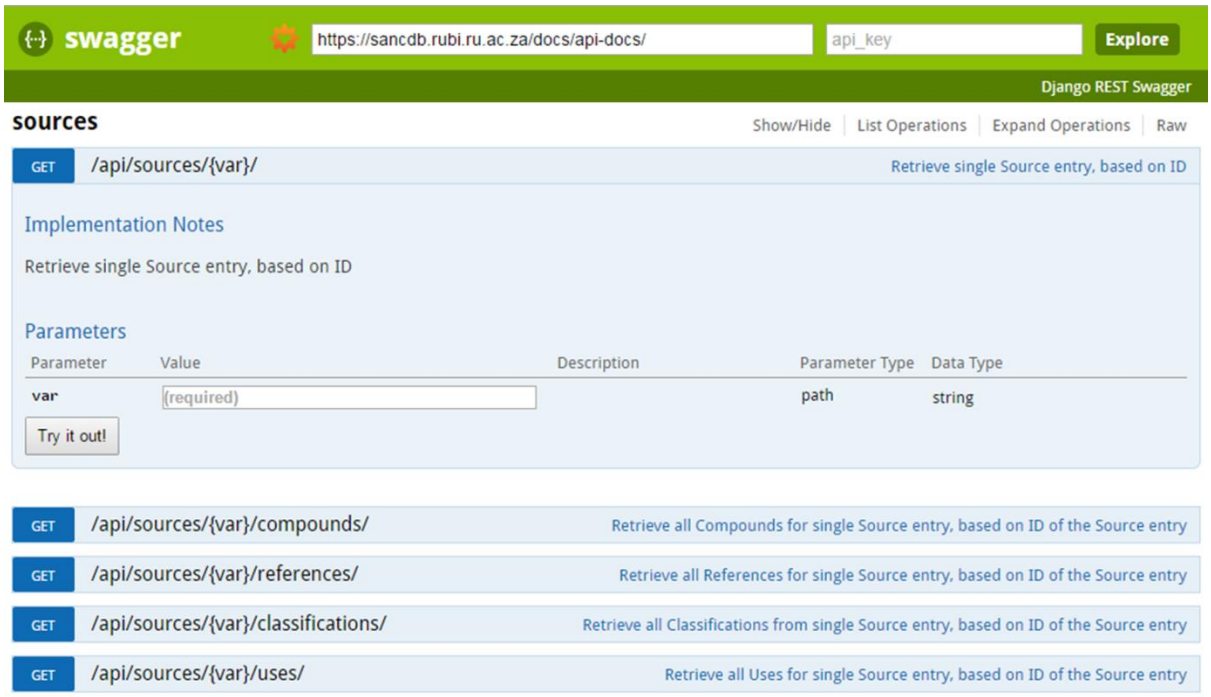


Figure 3.12: Django REST Swagger documentation page, showing retrieval options for web API.

3.3.5 Compound submission pipeline

The compound submission pipeline is a workflow, created by David Brown, which allows users to submit their own compounds to the database. It has been designed to mimic the upload process used by depositors and integrates scripts used to generate compound 3D coordinates, covert coordinate files into different formats, as well as perform minimisations. After compounds are submitted, they will be checked by curators from RUBi. If the compounds are from published literature, this process will also involve ensuring that all relevant information was extracted from these publications. Users with compounds that have

not been published may still submit their own compounds, but these will be linked to their users account. One of the goals of this project was to ensure that compound information was fully referenced, so as to be linked to an authoritative source. Until suitable literature is found, the user will be the authority on the compounds they submit.

3.3.6 Site documentation

The SANCDB website provides documentation for users that are new to the site. This page currently explains how to search for and download compounds. As the database expands to incorporate new information or new tools are integrated into the website, this page will be updated to describe and explain how to use these new features.

3.3.7 Integration of tools in future

Future developments to the site include enhancing the content of the database, as well as the integration of simple and useful tools. Currently the SANCDB website links to a web API and submission pipeline. There are several tools that are being developed, such as SMILES parser (Section 2.3.3) and the automated modelling tool (Chapter 6). The integration of these tools and their potential value to the site is discussed in Chapter 7.

3.4 Conclusion

SANCDB is an online interface to the South African natural products database established in Chapter 2. It is a user-friendly site that provides a number of different search options to find information collected about compounds in the database. Compounds can also be downloaded in a number of different formats, which will allow them to be further characterised or studied computationally. An API has also been added to provide an additional means to access the data housed within SANCDB and a compound upload screen has been incorporated to aid in the expansion of the database.

Section 2: Development of a homology modelling tool, based on protein structural studies

Chapter 4 Structural bioinformatics and its applications

4.1 Proteins, their structure and function

Proteins form an integral part of nearly all biological processes [164]. The structure of proteins have been shown to be responsible for their function [165]. The native state of a protein can be considered the overall spatial arrangement of its structure in which it is biologically active [164]. As such, a key goal of structural biology is to be able to determine the native structure of proteins in order to better study the mechanisms by which they function.

4.2 Structural genomics

Structural genomics projects aim to determine the structures of a large number of proteins [166]. This was brought about mostly by the success of genome projects producing a large quantity of sequence data, which has led to the establishment of a number of structural genomics projects in America, Asia and Europe [166]. This was aided by advances in both X-ray crystallography and NMR techniques, as well as the methodologies to express recombinant proteins [166]. The goal of structural genomics is to produce a protein structure for each sequence known for an organism. Proteins are, however, inherently more difficult to work with than DNA. They are far more diverse in their make up and vary greatly in physical properties and solubility. This has limited the number of proteins that have been produced as a pure, soluble sample with sufficient quality for X-ray crystallography and solution NMR [166].

4.3 Experimental methods of structure determination

4.3.1 X-ray crystallography

The bulk of the protein structures deposited in the PDB each year are solved using X-ray crystallography (Table 4.1 and Table 4.2). Since 1990, a number of improvements have been made to X-ray crystallography. Firstly, the processes involved in both the production and crystallisation of proteins have been automated. Other developments include the availability of more powerful X-ray sources and the advancement of the algorithms involved in collecting and processing X-ray diffraction data [167]. The use of robotics has also severely decreased the amount of pure protein required to produce crystals for structural determination X-ray crystallography. The first crystal structure, published in 1958, was that of myoglobin, solved at a 6 Å resolution [167, 168]. X-ray crystallography has come a long was since then, but can still be an incredibly difficult process, which requires high yields of pure protein crystals. This can prove challenging, as the process of crystallization is complex and not entirely understood. Often, protein samples fail to produce crystals of suitable quality, and sometimes crystals must be obtained through trial and error using a number of different approaches [167, 169]. The formation of protein crystals involves adding a precipitant to a protein solution. This can be achieved through a number of different methods; although the process of equilibrating the protein solution with the precipitant is becoming standardised, especially for automation [167]. Proteins are rarely produced directly from their source organisms for crystallography and are usually engineered to be expressed in *Escherichia coli* (*E. coli*) or certain eukaryotic cell lines [167]. The actual X-ray crystallography process involves exposing a crystalized protein sample to X-rays of a single wavelength that have been focused into a beam. This is repeated a number of times, while rotating the proteins crystal sample around a set axis. The X-rays are scattered by electrons within the protein structure and during this process the diffraction patterns of these X-rays are recorded [170]. This also

helps to determine the dimensions and orientation of the unit cell, which is defined as the smallest repeating unit that makes up a crystal [169]. The way the X-rays are diffracted, as well as the intensities of each individual diffracted X-ray, is based on the size and shape of the unit cell, as well as the position of each atom in the crystal [167]. As a result, modelling any given part of the unit cell requires the rest of the structure to be modelled as well, which is one way X-ray crystallography differs to other methods. The X-ray diffraction data yields an electron density map, which is processed to predict the positions of individual atoms and molecules within the crystal structure [167]. The electron density map is often refined to account for possible experimental errors [169]. The protein model produced from the electron density map is also refined based on energy minimisation and stereochemistry. An electron density map is calculated, based on the refined model and an R-factor is determined by comparing this to the experimental electron density data [169].

Table 4.1: Number of structures present in the PDB. The numbers of structures present in the PDB are listed by method used to solve the structure. Structures are divided into protein structures and "Other", which include nucleic acids, protein-nucleic acid complexes and heteroatoms. The information was obtained from the PDB on 01/06/2015.

Method	Proteins	Other	Total
X-RAY	90929	6155	97084
NMR	9611	1353	10964
EM	567	214	781
Hybrid	70	6	76
Other	165	23	188
Total	101342	7751	109093

4.3.2 NMR

One of the main advantages of NMR over other techniques is that the structure of proteins can be solved in solution [171]. Unlike X-ray crystallography, solution NMR works with protein in non-static poses and can be used to characterise protein dynamics, kinetics and thermodynamics [171]. This technique also solves structure at atomic resolution, but is

generally limited to proteins less than 25 kDa in size. For the purposes of protein structure determination, NMR uses the nuclear Overhauser effect (NOE) caused by nearby hydrogen atoms in the protein [172]. This allows different hydrogen atoms to be observed as long as they are within 5 Å of each other, even if they belong to different residues. Proteins are relatively large molecules and a single NMR run will produce thousands of peaks, even with small proteins [172]. Solving structures larger than 25 kDa compromises data quality unless higher-power magnets are used to improve resolution and prevent signal overlap [171]. Apart from signal overlap, there is a need to distinguish between peaks that are specific to the protein and those which are artefacts of the technique itself [173]. This is a very complex problem so the data is generally interpreted computationally. Protein NMR has seen some developments in recent years. Cell-penetrating peptides enables N_{15}-labelled proteins to be introduced into active cells for what is termed "in-cell NMR", which is carried out *in vivo* [174, 175]. Additionally, a number of different protocols have been developed to solve larger structures and the number of NMR studies with such proteins is increasing [171].

Table 4.2: Number of structures deposited in the PDB each year. This table includes all structures (protein, nucleic acid, etc.) deposited each since 2010, listed by method. The 2015 entries are up until 01/06/2015.

	X-Ray	NMR	EM	Total
2010	7194	522	54	7770
2011	7373	523	53	7949
2012	8187	537	65	8789
2013	8755	507	118	9380
2014	8912	554	191	9657
2015	2053	91	44	2188

4.3.3 Electron microscopy

Electron microscopy (EM) is related to light microscopy, but uses electrons rather than photons, thus allowing specimens to be observed at a molecular level [176]. Electron

scattering from contact with the biological sample results in the image that is formed and therefore the electron beam is kept in a vacuum to prevent scattering by gas molecules [177]. Image resolution quality could theoretically be comparable to X-ray crystallography (based on wavelength of electrons), but is primarily hindered by damage caused to samples by electrons [176]. Negative staining using heavy metal coating of samples prevents such damage, but does not provide structural information beneath this coat. Without negative staining, a less intense beam of electrons would need to be used, such that it does not damage the biological sample. Unfortunately, these samples consist mostly of low molecular weight elements which poorly diffract electrons, meaning this approach leads to a poor signal-to-noise ratio [176, 178]. By performing EM at very low temperatures (cryo-EM), radiation damage to samples can be reduced considerably, even when using an intense electron beam [176, 179]. Additionally, a large number of identical units are visualised and averaged as a supplementary means to improve resolution [176, 178]. EM is a technique that has been integrated into structural biology to study macromolecular assemblies that are not suited to study by NMR or X-ray crystallography [176].

4.4 *In silico* approaches

The prediction of the 3D structure of proteins can be achieved using either template-based modelling or so-called "free modelling" methods [180, 181]. In template-based modelling, the protein structure to be predicted is referred to as the "target". This is modelled based on one or more protein structures that have been determined experimentally, referred to as "templates". The template is usually evolutionarily related to the target protein, as in the case of homology modelling. However, in the absence of available homologous proteins with known structure, threading is used, which predicts secondary structure and protein folds to assign templates [180]. When homology searches and threading

methods are unable to identify suitable templates, free modelling methods are used which build models without the use of templates [181]. There is also a developing trend of overlap between homology modelling, threading and free modelling methods [180].

4.4.1 Homology modelling

Homology modelling works on the principle that structural features of a protein are far more conserved than its primary amino acid sequence. It has been determined that the structure of a protein may be conserved up to ten times more than its sequence [182]. For this reason, early attempts at homology modelling considered a template to be suitable for modelling only if it shared a high enough sequence identity with the target to be modelled [183–185]. Homology modelling is considered to be the most accurate form of *in silico* modelling. If an adequate homologous template is available, most homology models will have a root mean square deviation (RMSD) of 1 – 2 Å from the native structure of the target [180, 186]. This method is discussed in more detail in Chapter 6.

4.4.2 Threading

By its definition, threading is simply the process of aligning the sequence of a protein to the structure of a template [187]. Threading techniques rely on fold recognition, which uses force-fields and statistical potentials to select the most appropriate template for modelling, where there is no clear homologous structure available [188, 189]. The predicted folds can also be matched against a library of all known protein folds, to help identify the most likely fold for any given part of a sequence [190]. This is based on the concept that there are a limited number of structural folds proteins can adopt [187, 191]. The results obtained from threading are generally less accurate than those obtained through homology modelling methods. Models produced by threading will usually have an RMSD between 2 - 6 Å from the native structure [180, 192].

4.4.3 *Ab initio* methods

Free modelling methods were originally called "*ab initio*" methods, which aimed to predict the 3D structure of a protein, using nothing but the amino acid sequence. The ability to do this accurately is considered by many as the "holy grail" of structural biology [193]. Currently this is highly impractical as the purely *ab initio* methods available would take an inordinate amount of time to model even a small protein structure [194]. The computational time of these methods also increase exponentially with the length of the protein to be modelled. As a result, free modelling methods are now used, which combine these *ab initio* methods with knowledge-based approaches [181]. One of the most successful methods developed for free modelling is one which involves assembling fragments of known structures from the PDB. This was later incorporated into the program ROSETTA [195, 196], which is discussed in Chapter 6.

4.5 CASP

The Critical Assessment of Structure Prediction (CASP) initiative is a community-driven project which assesses different protein modelling methods [197]. One of the aims of CASP is to identify the pitfalls of modelling techniques and determine which challenges need to be addressed in order to improve the accuracy of theoretical models. CASP runs every two years and participants are given targets that have not yet been solved and are also given a wide range of proteins to model to account for methods with specific biases. The evaluation of structures is also done using a number of different techniques for the same reason. CASP is seen as a competition, as different groups will model the same target proteins. The results and models are all freely available afterwards so that the data may be used by anyone [197]. CASP targets are ranked by difficulty and groups may enter based on the method of structure

prediction, such as comparative modelling, free modelling, automated servers or just prediction of model quality.

4.6 Uses of protein structural data

One of the purposes of structural biology is to understand how organisms function at a molecular level [198]. Protein structural data can be used for drug discovery purposes; both for the identification of lead compounds and further development into more potent drugs [199]. As such, many efforts have gone into solving the structures of potential drug targets [199]. Structural biology has also helped characterise protein-protein interactions, the assembly of multiple-protein complexes and helped to elucidate mechanisms involved in biochemical pathways [200]. Functional predictions have been made about proteins under the assumption that proteins with similar structures should have similar functions. This is not always the case, however, as some proteins with similar folds will have different functions, even if they have similar active sites [198]. Structural data can be studied to understand known functional differences in these cases [201]. Problems faced with protein expression and crystallisation may also be overcome through analysis of available structural data [198].

4.7 Successful use of computational protein modelling

Although many homology modelling efforts are aimed at ligand docking for drug discovery, this method is also used to characterise proteins whose structures have not yet been solved experimentally. A good example of this is the modelling of purine riboswitches by Sharma *et al*. [202]. The group modelled the active site of both adenine and guanine riboswitches allowing them to use various forms of binding interaction analysis and energy analysis. This granted a mechanistic insight into the function of these proteins and how they interact with their metabolites, as well as the mechanisms involved in substrate specificity. Advances in

predictive techniques has provided more structures to be used to guide drug discovery efforts [203]. Homology modelling and virtual screening was used to successfully identify and develop compounds that were highly effective agonists of muscarinic acetylcholine receptor [204]. The final compound produced this way showed indications of cognitive enhancement when tested in animal models. Homology modelling and ligand docking has been used to target the P-protein involved in melanin production, for cosmetic purposes [205]. Using these methods to screen a molecular library of potential inhibitors resulted in the identification of five compounds that, when tested *in vitro*, resulted in reduced melanin production and minimal toxicity. Recently, structural studies were performed using computational models of cytochrome B of *Plasmodium falciparum* (*P. falciparum*) [206]. This was done to investigate the rapid resistance of *P. falciparum* to the drug, atovaquone, which targets this protein. By combining structural modelling with both ligand docking and molecular dynamics, Akhoon and co-workers demonstrated than even a single residue change to the active site of cytochrome B, led to atovaquone binding elsewhere on the protein, rendering the drug ineffective.

4.8 Research aim

Structural bioinformatics provides a means to study proteins in such a way as to gain a mechanistic insight into how they function. This section of the thesis initially focused on performing structural characterisation of *P. falciparum* heat shock proteins, with the goal of characterising their interactions with prospective cochaperones. This involved performing homology modelling a number of times, often remodelling using different parameters until suitable models were obtained. This gave rise to the development of a homology modelling tool, which can perform the various steps involved in homology modelling from identification of a suitable template, to assessment of the models produced.

Chapter 5 *P. falciparum* structural studies

This chapter is centred around two of my published works and therefore describes two different studies. The content of these publications is described here in more detail. The first [207], focuses on one of the *P. falciparum* 70 kDa heat shock protein (PfHsp70s), namely PHsp70-x. This is one of the six PfHsp70s, and the only one that does not have an orthologues in other plasmodial species. Potential interaction sites between PfHsp70-x and exported *P. falciparum* and host erythrocyte J proteins were interrogated by modelling and protein-protein docking. Docking results indicated that interactions between PfHsp70-x and each of the Hsp40s tested seem equally likely, suggesting that the J domain may not provide the specificity in the formation of unique Hsp70-Hsp40 complexes, but that the specificity might be provided by other domains of Hsp40s. By studying different structural conformations of PfHsp70-x, it was shown that Hsp40s can only bind when PfHsp70-x is in a certain conformation. The second study [208] concerns the heat shock organizing protein (Hop), which is important in modulating the activity and co-interaction of Hsp70 and Hsp90. Recent research has suggested that *Plasmodium falciparum* Hop (PfHop), PfHsp70 and PfHsp90 form a complex in the trophozoite infective stage. There has been almost no computational research on malarial Hop in complex with other malarial Hsps. Homology modelling was used to study the interactions of PfHop and human Hop (HsHop) with their own cytosolic Hsp90 and Hsp70 C-terminal peptide partners. The results indicated that these interactions at the concave TPR sites of Hop are highly conserved between both *P. falciparum* and host proteins. Analysis of additional binding sites between the convex sites of Hop TPR2 motif and the Hsp90 middle domain showed that these interactions are distinctly less conserved between human and the malaria parasite. The low conservation of these convex sites indicates that their potential for malarial inhibitor design may be more viable than the concave sites, which have been the focus of previous efforts.

5.1 Introduction

5.1.1 Malaria

Malaria is the most devastating parasitic disease of the current age. The World Health Organisation (WHO) reported that an estimated 3.2 billion people are at risk of malaria infection, with approximately 198 million cases occurring annually [209]. Of the 584 000 estimated malaria deaths in 2013, 90% occurred in Africa and 78% of the total deaths were children under the age of five [209]. There is also a great economic burden associated with malaria. Currently funding more malaria efforts is at 2.7 billion US dollars, which is just over half the amount required to control the disease [209]. The disease is caused by parasites of the genus *Plasmodium* and spread by an insect vector; the female *Anopheles* mosquito [210]. There are five species within the *Plasmodium* genus that cause malaria in humans – *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* and most recently *P. knowlesi.* Most focus on *P. falciparum* and *P. vivax*, the latter of which produces far milder symptoms, but is the most wide-spread form of malaria and can exist at a larger range of environmental conditions and geographic locations. The most lethal species is *P. falciparum*, which accounts for the majority of deaths that occur, especially in Africa [209].

5.1.1.1 Life cycle of *P. falciparum*

The life cycle of *P. falciparum* involves stages in both vertebrate human host and invertebrate mosquito hosts, both with vastly different internal environments [211]. There are a number of stages in the development and reproduction of *P. falciparum*, illustrated in Figure 5.1. These include the sporozoite, merozoite, trophozoites and gametocyte stages; although a ring stage is also said to occur between merozoite and trophozoite stages [210]. The parasite enters the human host when the mosquito feeds and releases

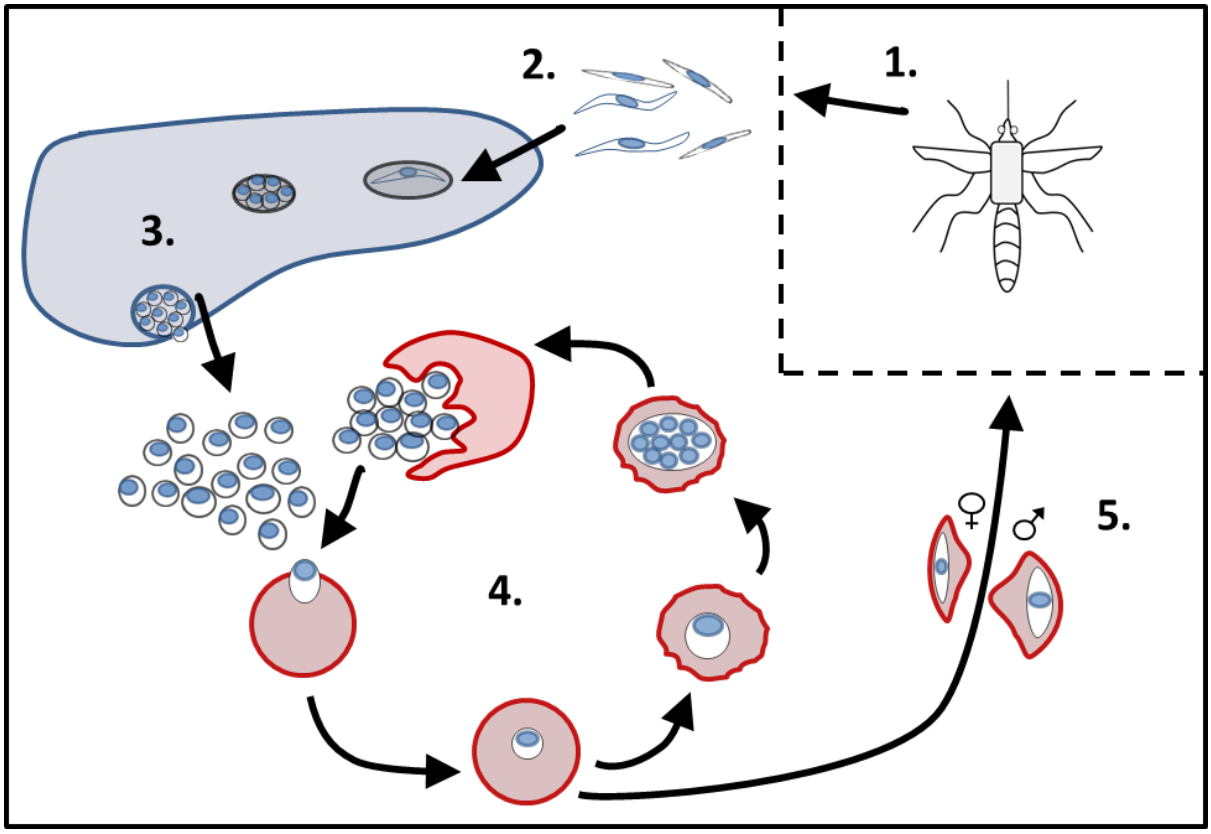


Figure 5.1: Life cycle of *P. falciparum*. Illustrated are the stages which take place inside the human host, discussed in text. The stages are labelled as follows: 1) Mosquito feeding, releasing sporozoites into the host's bloodstream; 2) Sporozoites invading the hepatocytes; 3) Release of merozoites from the liver; 4) Erythrocytic stages of the malaria lifecycle; 5) Gametocytes taken up by the mosquito during feeding. Adapted from Fujioka and Aikawa (2002) **[210]**.

thousands of sporozoites into the bloodstream, which invade liver cells (hepatocytes). *P. falciparum* lacks dormant phase in the liver, seen in other *Plasmodium* species, such as *P. vivax*, which can stay in the liver up to a number of years, causing relapse of the disease [210]. Thousands of merozoites are released from the liver and invade red blood cells (RBCs). The trophozoite stage includes remodelling of the RBC and modification of its surface. The contents of RBC cytoplasm are also digested in food vacuole. This includes haemoglobin, which is degraded to provide nutrients for the growth of the parasite [212]. Following this is asexual reproduction, involving the production of merozoites, which are then released into the blood after RBC rupture [211]. The gametocytes are involved in sexual reproduction, which only occurs in the mosquito host [211]. They exist in two structurally

distinct forms, simply named male and female and are taken up by the mosquito when it feeds. A thorough understanding of the processes that occur during the life cycle of the malarial parasites are considered to be important for developing drugs which target each stage [211].

5.1.1.2 Treatment strategies and drug resistance

Currently, the most effective treatment strategies include preventative measures, chemoprevention and diagnostic work [209]. The use of insect-treated nets (ITNs) is predicted to reduce the number of malaria infections by half [213]. ITNs are a favoured means of prevention, since they may also hinder other diseases that are transmitted by insect vectors [214]. Chemopreventive measures are part of an intermittent preventive treatment (IPT) approach, which involves issuing antimalarial treatment at predetermined times to prevent malaria, rather than treat it [215]. Chemopreventive therapies, such as sulfadoxine-pyrimethamine [216] and mefloquine [217] have been shown to be effective [218, 219]. The early stages of malaria are referred to as uncomplicated malaria, which develops into severe malaria over time. Severe malaria is life-threatening if untreated [219]. The current treatment for uncomplicated malaria, recommended by the WHO is artemisinin-based combination therapies (ACTs) [209, 220]. These combine artemisinin derivatives with longer-lasting drugs and have been used to successfully treat malaria [219]. *P. falciparum* has developed resistance to most antimalarial drugs available, including amodiaquine, chloroquine, mefloquine, quinine, and sulfadoxine-pyrimethamine [221]. More recently, multiple cases of artemisinin resistance have also been reported [209]. These reports of *P. falciparum* developing drug resistance to an increasingly large number of antimalarial drugs highlights a need to identify new drug targets to combat the parasite.

5.1.2 Heat shock proteins

A number of studies have identified the heat shock proteins (Hsps) of *P. falciparum* as prospective drug targets [222–228]. Hsps are a group of molecular chaperones, which are proteins that assist with the correct folding of other proteins and prevent protein aggregation [229]; although Hsps may also perform other functions [230]. Hsps are up-regulated during times of physiological stress, including temperature increase, hypoxia, the presence of heavy metals and radiation, as well as glucose deprivation [231]. In *P. falciparum*, Hsps perform essential functions as part of the heat shock response of the parasite, helping it survive the temperature increase of more than 10°C when moving from the mosquito vector to its human host, as well as during recurrent fevers experienced by patients with malaria [228]. Hsps have also been implicated in host RBC remodelling [228] and regulation of actin polymerization [232], which is linked to the actin-myosin motor system, thought to facilitate RBC invasion by merozoites [233]. Subjeck *et al*. [234] reported that the proteins most greatly up-regulated during times of stress are the 70 kDa, 90 kDa and 110 kDa Hsps (Hsp70, Hsp90 and Hsp110, respectively).

5.1.2.1 Hsp70s

In addition to performing the roles associated with molecular chaperones, Hsp70s assist with the folding of newly translated proteins, the translocation proteins across membranes, the disassembling of protein complexes, as well as the degradation of unstable proteins [235]. These proteins bind and release peptide substrates in an ATP-dependant cycle, assisted by Hsp40 cochaperones [236]. The domain structures of Hsp70, along with Hsp90, J proteins and the Hsp70/Hsp90 organising protein (Hop), is illustrated in Figure 5.2.

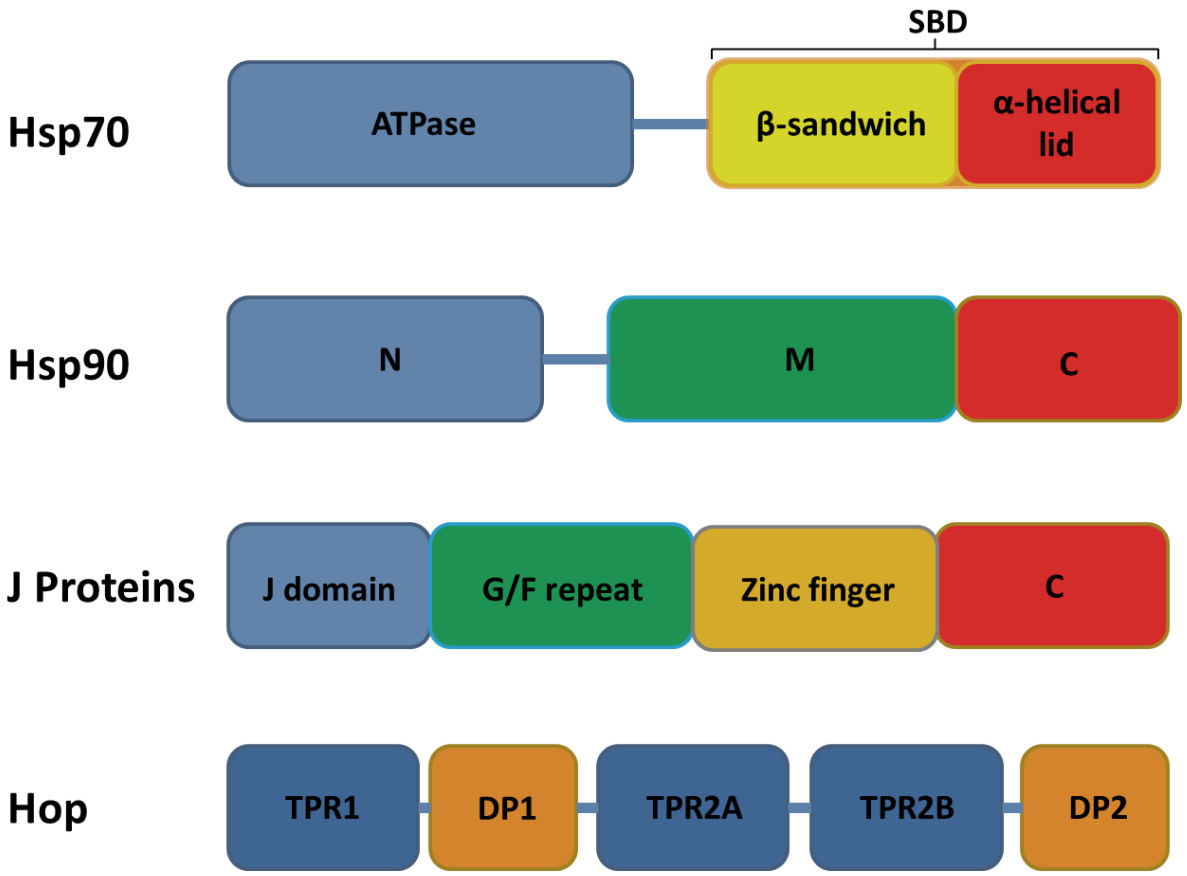


Figure 5.2: Domain structures of Hsp70, Hsp90, J proteins and Hop. The domains that make up each protein are shown, with their relative positions in each protein. Each is discussed in more detail in text.

5.1.2.1.1 Structural features of Hsp70

Hsp70s consist of two major structural domains [237]. Firstly, the 45kDa N-terminal domain, also called the nucleotide binding domain (NBD) or ATPase domain; and the 25 kDa C-terminal domain, otherwise known as the peptide-binding domain or substrate-binding domain (SBD). These two domains are connected by a flexible linker region, which facilitates communication between the ATPase domain and SBD [238, 239]. The ATPase domain and SBDs both regulate each others' functions [240]. The SBD is further subdivided into a β-sandwich subdomain, which contains the substrate-binding pocket, and the α-helical lid subdomain [241, 242]. This lid region is not directly involved in substrate binding, but

does close over the bound substrate and has been found to stabilize substrate binding [243, 244].

5.1.2.1.2 ATPase cycle of Hsp70

During the ATPase cycle of Hsp70 Figure 5.3, the protein undergoes transitions between an ATP-bound state, in which it has a relatively low affinity for peptide substrates, and an ADP-bound state, in which it has a high affinity for peptide substrates [235]. In ATP-bound state it can rapidly cycle between substrates and it is in this state that Hsp70 initially binds its peptide substrates. ATP hydrolysis is then stimulated by a J protein cochaperone, causing Hsp70 to strongly bind the peptide substrate [245, 246]. There is also evidence to suggest that through its own peptide binding activity, the J protein recruits misfolded peptide substrates to be bound by Hsp70 [247]. Another cochaperone, the Hsp70-interacting protein (Hip), stabilizes the ADP-bound form of Hsp70 by preventing release of ADP [248]. Regeneration of the ATP-bound state of Hsp70 is facilitated by a nucleotide exchange factor (NEF), which binds to the ATPase domain of Hsp70 and lowers its affinity for ADP [249, 250]. The ADP from Hsp70 is exchanged for the ATP from the NEF and, in an ATP-bound state, Hsp70 releases the bound peptide from its SBD and is able to bind another misfolded protein.

5.1.2.1.3 P. falciparum Hsp70s

A total of six Hsp70 genes have been identified in *P. falciparum* [251], given the names PfHsp70-1, PfHsp70-2, PfHsp70-3, PfHsp70-x, PfHsp70-y and PfHsp70-z [230]. Of these, PfHsp70-1 has by far been studied in the greatest detail.

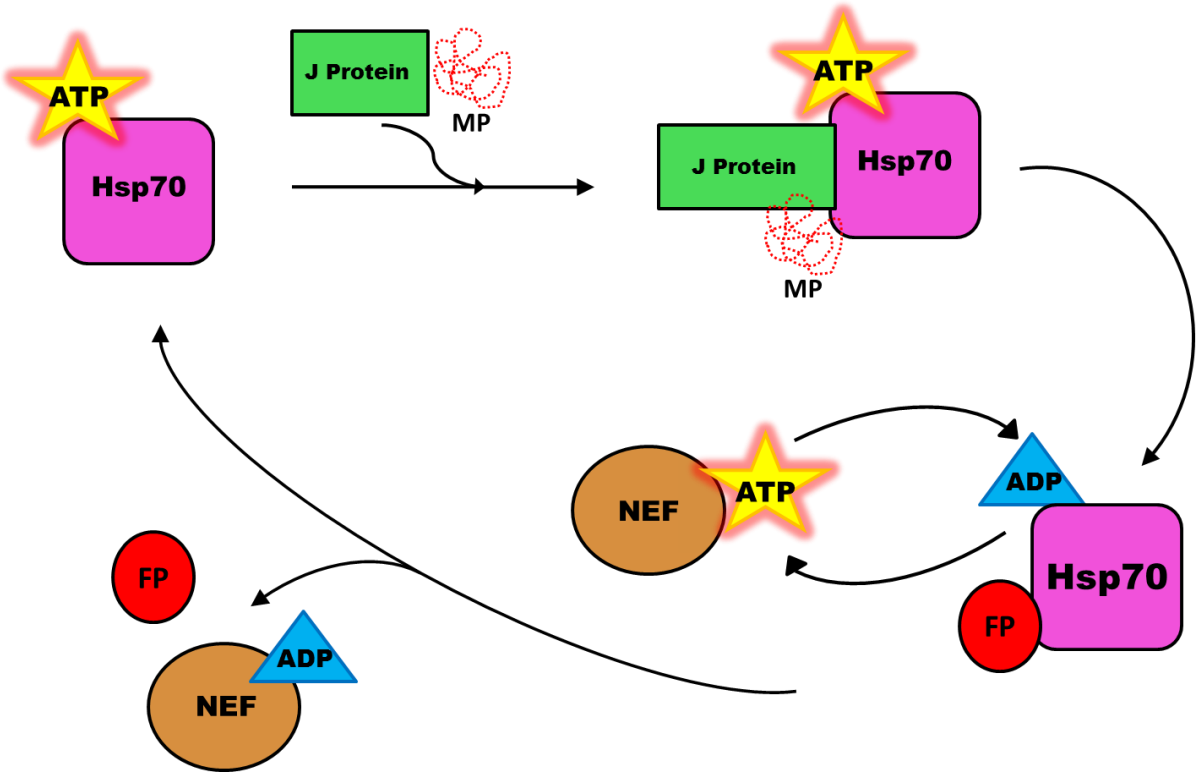


Figure 5.3: ATPase cycle of Hsp70. The ATP-dependent binding of a misfolded peptide (MP) substrate and release of a correctly-folded protein (FP) by Hsp70 is indicated, as described in text.

PfHsp70-1 is located in the cytosol of *P. falciparum* and expressed in all blood stages of the parasite's life cycle [252]. It has been shown to be capable of reversing the thermosensitivity of an *E. coli* strain with a partially defective DnaK [253]. A chimeric Hsp70 protein, composing of the ATPase domain of *E. coli* DnaK and the SBD of PfHsp70-1 was shown to be functional when expressed in a thermosensitive *E. coli* strain [253]. PfHsp70-1 was also able to reproduce some of the functionality of yeast Hsp70s, Ssa1 and Ssa2, when expressed in strains lacking these proteins [254]. The importance of the linker region of PfHsp70-1 has also been established, as mutations in this region rendered the protein non-functional [253]. PfHsp70-1 can also suppress thermally-induced malate dehydrogenase aggregation, as well as reactivate denatured glucose-6-phosphate dehydrogenase [255] PfHsp70-1 may also facilitate the transport of proteins to the apicoplast of the malarial parasite [256, 257]. PfHsp70-1

is also able to interact with the heat shock binding protein of *P. falciparum* (PfHSBP), indicating that it may play a role in the heat shock response mechanism of the parasite [258].

The other cytosolic Hsp70 candidate of *P. falciparum*, PfHsp70-x [251] has only recently received any research attention. The expression of the protein was first confirmed in 2009 by mass spectrometry of the proteome of clinical isolates of the parasite [259]. PfHsp70-x had never been detected in laboratory strains, which suggested that it was only expressed pathogenic stages of the parasite's life cycle. PfHsp70-x has been shown to be partly exported to the host erythrocyte, and partly present in the PV [260, 261]. Another interesting aspect of PfHsp70-x is the carboxy-terminal EEVN motif, since an EEVD motif is characteristic of cytosolic eukaryotic Hsp70s [230].

Like PfHsp70-1, PfHsp70-2 is expressed in all blood stages of the parasite's life cycle [252, 262, 263]. This protein localises in the endoplasmic reticulum (ER) of *P. falciparum*. There has been little work done to characterise PfHsp70-3, which is thought to reside in the mitochondria of the parasite, due to similarity to other eukaryotic mitochondrial Hsp70s [264]. There is evidence to suggest that this protein may localise in the PV of the *P. falciparum* [265] and also may be targeted to the apicoplast [257]. PfHsp70-y and PfHsp70-z, appear to be Hsp110 homologues [230]. Although these are structurally related to Hsp70s, they are part of a different and distinct sub-family of proteins [266]. PfHsp70-y and PfHsp70-z are predicted to localise in the ER and cytoplasm, respectively [251].

5.1.2.2 Hsp90s

Hsp90s are not required for folding of most proteins, but rather for the final maturation of specific target proteins, referred to as "clients" [267–269]. Hsp90 acts on proteins whose native states are more difficult to be achieved and, as such, during heat stress, its range of client proteins increase [267]. This protein also does not prevent denaturation of other

proteins, but will facilitate refolding [267]. In addition to protein folding, the functions of Hsp90 include a role in signal transduction, intracellular transport, and protein degradation [268]. The protein has been implicated in the assembly of numerous multi-protein complexes and is involved in many cellular pathways [270].

Like Hsp70, the function of Hsp90 is linked to an ATP-dependant cycle (Figure 5.4), involving conformational changes [271]. However, Hsp90 acts as a homodimer with a parallel arrangement [268, 269]. Its activity occurs through a nucleotide-bound molecular clamp mechanism, wherein the homodimer goes through open and closed conformation during its ATPase cycle [269, 272, 273]. The monomeric structure of Hsp90 consists of an N-terminal (N) domain, a middle (M) domain and a C-terminal (C) domain [269, 274], as indicated in Figure 5.2. The N domain is the site of the ATP-binding pocket and therefore responsible for the ATPase activity of Hsp90 [275]. It also contains a number of residues that form a lid, which closes over the ATP-binding site when ATP is bound [269]. The N and M domains are connected by a flexible linker region that is required for Hsp90 functioning and believed to facilitate inter-domain communication [274, 276]. The M domain contains a catalytic loop believed to recognise the gamma-phosphate of ATP and initiate ATP hydrolysis [274]. The M domain is also believed to interact with client proteins [274]. Hsp90s dimerize at the C domain, as well as a portion of the M domain [269, 277]. There are also interactions between N domains in ATP-bound state [269, 273].

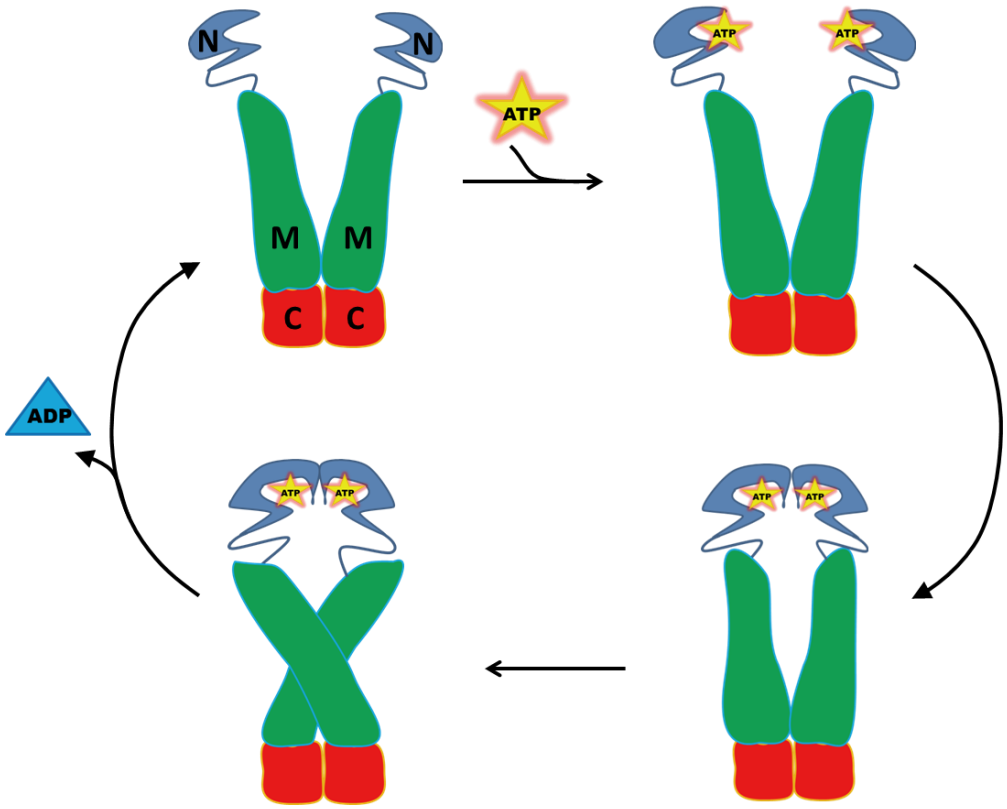


Figure 5.4: ATPase cycle of Hsp90. Hsp90 is shown as a dimer, with each domain coloured differently, as labelled in the top left structure. The process of the ATPase cycle is described in text.

5.1.2.2.1 Cochaperones of Hsp90

In prokaryotes, evidences suggests that Hsp90s (HtpG) act without the need of cochaperones. In eukaryotes, however, over 20 different cochaperones have been implicated in regulating the function of Hsp90 [278]. These are reviewed in detail by Li *et al*. [278]. Examples include cochaperone p23 (Sba1 in yeast) which stabilizes the closed conformation of the Hsp90 homodimer by inhibiting the hydrolysis of ATP [269, 279]. This allows client protein maturation by preventing its release. In contrast to this, another cochaperone, known as the activation of Hsp90 ATPase protein 1 (Aha1) stimulates ATPase activity of Hsp90 [280, 281]. An interesting cochaperone is the Hop, which facilitates transfer of protein substrates from Hsp70 to Hsp90 [282]. Hop is one of the proteins focused on in this chapter and is discussed in more detail in Section 5.1.2.4.

5.1.2.2.2 *P. falciparum* Hsp90s

In *P. falciparum*, there is only one cytosolic Hsp90 (PfHsp90), as opposed to the two cytosolic Hsp90s found in humans [227, 283]. Like its human counterparts, PfHsp90 interacts with client proteins involved in cellular pathways, as well as kinases and transcription factors [227]. The mechanisms of PfHsp90 function have also been implicated with enhancing the parasite's pathogenesis during febrile episodes experienced during malarial infections [284]. PfHsp90 may play a role in development of resistance to antimalarial drugs [283]. PfHsp90 has also been shown to be essential to the parasite's survival and is a promising antimalarial drug target [223, 225, 227].

5.1.2.3 J proteins

J proteins are traditionally called Hsp40s in eukaryote organisms, but the molecular weight of these proteins can often vary quite significantly from 40 kDa. For this reason, they are more recently being referred to collectively as J proteins [285, 286]. As mentioned above, these cochaperones stimulate the ATPase activity of Hsp70s [287]. J proteins are incredibly diverse and have different domain structures [288]. As such their only definitive structural feature is the presence of a conserved J domain through which it interacts with Hsp70 [286, 288–291]. The J domain is approximately 70 residues in length and contains a Histidine-Proline-Aspartic acid (HPD) motif [288]. The HPD motif is required to stimulate ATPase activity of Hsp70 [292]. There are four proposed types of J proteins, based on the presence and architecture of specific structural features, which are simply referred to as types I-IV [288, 293]. Type I J proteins, based on DnaJ from *E. coli*, contain a J domain, followed by a flexible glycine- and phenylalanine- (G/F) rich region and a Cysteine-rich Zinc finger. Type II only contains the J domain and the G/F-rich loop and type III only contains the J domain, located anywhere within the protein [288]. Type IV J proteins also only have a J domain, but have a variation on the conserved HPD-motif [293].

5.1.2.3.1 *P. falciparum* J proteins

The *P. falciparum* genome is currently believed to encode at least 49 different J proteins, 19 of which are exported [286]. There are two Type I J proteins found in the parasite [293]. One of these, PfJ1 (NP_702750.1), has been found to be up regulated, at an mRNA level, during heat shock [294], but has been shown to localise in the apicoplast [295]. The other, now termed PfHsp40 (NP_702248.1), is also up regulated during heat shock and has been confirmed to localise in the cytosol, as well as moderately stimulate the ATPase activity of PfHsp70-1 [296]. This protein was also found to suppress protein aggregation when used alone, as well as enhance the ability of PfHsp70-1to suppress protein aggregation [296]. Additionally, PfJ4, a known *P. falciparum* Hsp40, which localises in the cytoplasm and nucleus of the parasite, has been shown to associate with PfHsp70-1 [297]. Recently, two type II J proteins, PFE0055c and PFA0660w (PfA and PfE, respectively), have been found to be exported and localise along with PfHsp70-x in structures called "J dots" [260]. Another *P. falciparum* J (PfJ) protein, PFB0090c (PfB), has also been predicted to be exported to the erythrocyte cytosol [293].

5.1.2.4 Hop

The Hop cochaperone (otherwise known as p60; or Sti1 in yeast) has not been as well-studied as Hsp70 or Hsp90, but plays an important role in facilitating the transfer of client proteins from Hsp70 to Hsp90 [282, 298, 299]. When client protein is bound, Hop inhibits ATPase activity of Hsp90 [300]. PfHsp70-1 and PfHsp90 have been shown to form part of the same multi-chaperone complex, as seen in other systems [223]. Interestingly, evidence suggests that host chaperones Hsp70, Hsp90 and Hop are recruited by the parasite in detergent-resistant membrane-bound complexes [301]. It has recently been suggested that the Hsp70-Hop-Hsp90 complex forms within the host erythrocyte during the trophozoite stage of the *P. falciparum* life cycle [302].

5.1.2.4.1 Structure and function of Hop

Hop consists of three tetratricopeptide repeat (TPR) domains and two DP domains (Figure 5.2). Currently there is no experimentally-determined full-length structure for Hop [298]. Each TPR domain has a concave and convex surface [303]. There have been various studies which focus on the mechanisms of both the Hop:Hsp70 and Hop:Hsp90 interaction [298, 299, 304, 305]. What has been determined is that the concave surfaces of the TPR1 and TPR2B domains bind the C-terminal EEVD peptide of Hsp70 [298, 299, 304, 305] and the concave surface of the TPR2A domain C-terminal MEEVD peptide of Hsp90 [305–308]. Richter *et al*. [309] determined that the binding energy of the concave interactions between Hop and Hsp90 did not account for total affinity between these two proteins. Schmid *et al*. [298] suggest a conformation for this transfer, shown in Figure 5.5.

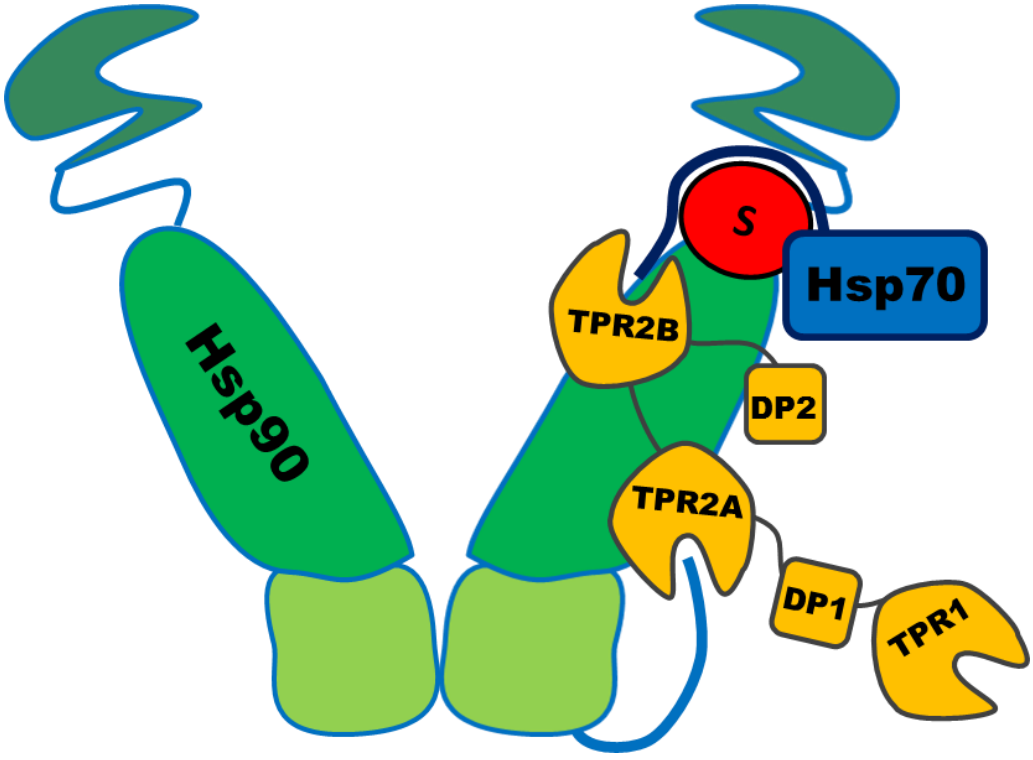


Figure 5.5: Hop-mediated substrate transfer from Hsp70 to Hsp90, as proposed by Schmid *et al*. [298]**.** The substrate (S) transferred is coloured red, while Hop is coloured in orange, with its individual domains labelled. This conformation allows the concave surfaces of TPR2A and TPR2B to bind the C-terminal peptides of Hsp90 and Hsp70, respectively, while the convex surfaces of these domains interact with the Hsp90 M domain.

5.1.2.5 Drugs targeting Hsps

Geldanamycin is a drug that has been shown to inhibit the function of Hsp90 [223, 225, 310, 311]. It has also been shown to be an effective antimalarial against chloroquine-resistant strains and can therefore be used in combination with this drug [225]. Geldanamycin and its many derivatives [312] act by competitively binding to the Hsp90 nucleotide binding cavity, inhibiting its ATPase activity [275]. There have been a number of other Hsp90 inhibitors, but these have been designed to combat cancer [312–314]. Shahinas *et al*. [315] screened over 4000 FDA-approved drugs for Hsp90 ATPase inhibitors and found compounds 2-Amino-3-phosphono propionic acid, harmine and acrisorcin, which showed potent antimalarial activity, as well as synergistic effects with chloroquine, as seen with Geldanamycin [225]. The majority of Hsp90 drugs target the ATPase activity of the molecular chaperone [312, 316]. Other inhibitors, such as novobiocin, which targets the C-terminal of Hsp90 [316, 317], and celastrol, gedunin and H2-gamendazole, which target interactions with cochaperones [317]. Cortajarena *et al*. [318] designed peptides that mimic the interaction between the Hop TPR2A domain and Hsp90, inhibiting Hsp90 activity. Unlike other Hsp90 inhibitors, this did not result in increased expression of Hsp70, which is a perceived advantage of targeting the Hsp90:Hop interaction.

There are fewer Hsp70 drugs that have been developed, though this may be an emerging drug target [319]. The first inhibitor of Hsp70 was 15-deoxyspergualin (DSG), which enhances the ATPase activity of Hsp70 [320, 321]. Compounds related to DSG and their derivatives have displayed Hsp70 inhibition and are being developed in an effort to identify potential anticancer drugs [321]. Anticancer research efforts have found a number of different Hsp70 inhibitors that target ATPase and substrate-binding activity, as well as interactions with cochaperones, such as Hip and Hop [314, 321]. Studies have also been performed to identify Hsp70 inhibitors that have potential as antimalarial agents. Chiang *et al*. [224] identified nine

pyrimidinone Hsp70 inhibitors that displayed activity against *P. falciparum*, including the compound DMT2264, which had stronger inhibitory effects towards PfHsp70-1 than on both yeast and human homologs, Ssa1 and HsHsp70. Cockburn [322] identified malonganenone A, which specifically inhibited both PfHsp70-1 and PfHsp70-x and not the human Hsp70, Hsp70A1A. This compound was also found to display antimalarial activity, while displaying relatively low toxicity towards mammalian cells. Pesce *et al*. [319] suggest that targeting Hsp90, along with cochaperones, such as Hop could be a good strategy and chances of developing resistance would be lower. Also raise concerns about targeting PfHsp90 and not hostHsp90. Also suggest targeting Hsp40:Hsp70 system could be very effective.

5.1.3 Proposed work

This chapter initially covers work that following on from the structural characterisation of PfHsp70 proteins, performed during my masters [323]. Specifically, this involves exploring the potential interactions between PfHsp70-x and other exported malarial J proteins and host J proteins, using both homology modelling and protein-protein docking. Also reported is the structural work surrounding the comparative characterisation of the Hsp90:Hop convex interaction interface of human, yeast and *P. falciparum* proteins and comparing these interaction sites to the concave interaction interface, which is the targeted region of the Hsp90:Hop interaction in current drug therapies.

5.2 Methodology

Unless otherwise stated , all programs were run using default parameters.

5.2.1 Sequence acquisition

The sequences of PfHop (PF3D7_1434300), PfHsp70-1 (PF3D7_0818900), PfHsp70-x (PF3D7_0831700), PfHsp90 (PF3D7_0708400), PfA (PF3D7_0113700), PfB (PF3D7_0201800) and PfE (PF3D7_0501100.1) were downloaded form PlasmoDB v9.3 [324]. Additionally, HsHsp90 alpha (accession number: NP_005339.3) and HsHsp90 beta (accession number: NP_031381.2) and HsHsp70-1A (accession number: NP_005337.2) were obtained from the NCBI's Protein database [325]. As well as sequences for two type I human Hsp40s, HsDnaJA1 (NP_001530.1) and HsDnaJA2 (NP_005871.1), two Type II human Hsp40s, HsDnaJB1 (NP_006136.1) and HsDnaJB4 (NP_008965.2), and two Type III human Hsp40s HsDnaJC5 (NP_079495.1) and HsDnaJC13 (NP_056083.3).

5.2.2 Homology modelling

5.2.2.1 Acquisition of templates

Templates for modelling were identified using HHPred [326], with the exception of the coordinate file for the *Saccharomyces cerevisiae* (*S. cerevisiae*) Hop:Hsp90 (ScHop:ScHsp90) convex complex, which was kindly supplied by Schmid *et al*. [298]. This structure contained the ScHopTPR2AB domain complexed with ScHsp90 M and C domains (ScHopTPR2AB-ScHsp90MC). The authors calculated this by docking an ScHopTPR2 fragment (PDB ID: 3UQ3) to a low-resolution spin-labelled model ScHsp90 M and C domain fragment. During the preparation of this coordinate file, Hsp90 residues S-411, S-422 and S-184 with CXM residues, which is a Cysteine residue with an additional proxyl group [298]. Two different templates were prepared: Firstly, the CYS-template, where these three residues

were converted to Cysteine residues, by simply removing the proxyl group in the coordinate file; and secondly, a SER-template, where each of these residues in the CYS-template was converted to Serine, using MODELLER's mutate-function.

5.2.2.2 Modelling

Alignments of template sequences to target sequences were done using PROMALS3D [327] and corrected manually by inspection, if necessary. Modelling was performed using MODELLER [328], with slow refinement. For each protein or complex, 100 models were built. The best three models were selected based on their calculated DOPE Z scores. These models were further evaluated using MetaMQAPII [329]. Final evaluations were performed using MetaMQAPII, Verify3D [330] and ProSA [331, 332].

5.2.2.2.1 Conformations of full-length PfHsp70-x

The full length structure of PfHsp70-x was modelled in two different conformations. Firstly, in its ATP-bound conformation (i.e. before hydrolysis of ATP), using bovine Hsc70 (PDB ID 1YUW, chain A) [333] as a template. Secondly in its substrate- and ADP-bound state (i.e. after hydrolysis of ATP), using *E.* coli DnaK (PDB ID 2KHO, chain A) [334] as a template.

5.2.2.2.2 PfHsp70-x complexed with J proteins

The template used for modelling the Hsp70:J domain interaction was 2QWO [291]. This contained ATPase domain of bovine Hsc70 disulphide-cross-linked to a bovine Type III J protein, auxillin. Due to the low sequence identity between auxillin and the different Hsp40s being studied, a hybrid modelling approach was used [335] to model these complexes. First, each Hsp40 J domain was modelled separately as monomers. Templates with the following PDB IDs were used for modelling each J protein: PfA - 1HDJ, PfB - 2CTR, PfE - 2CTP, HsDnaJA1 - 2O37, HsDnaJA2 - 1HDJ, HsDnaJB1 – 1HDJ, HsDnaJB4 – 1HDJ, HsDnaJC5 –

2CTW and HsDnaJC13 – 2CTW. Each J domain model was superimposed with auxillin of template structure to replace auxillin, using PyMOL [336]. These hybrid structures were used as templates for modelling PfHsp70-x and each J protein complexes. The resulting models underwent residue repacking and were then put through 100 cycles of energy minimisation, both done using PyRosetta [337].

5.2.2.2.3 Hsp90 with the convex surface of Hop

The ScHopTPR2AB-ScHsp90MC template was used to model the convex interaction interface of Hop with Hsp90 for the following combinations of proteins: 1) PfHsp70-1 with PfHop; 2) HsHsp90 alpha with HsHop; and 3) HsHsp90 beta with HsHop.

5.2.2.2.4 Hsp90 with the concave surface of Hop

Interactions between the concave surfaces of Hop and the C-terminal peptides of Hsp70 and Hsp90 were modelled using templates 1ELW, 3UQ3 and 3UPV. These templates contained the following: 1) PDB ID 1ELW: Human HopTPR1 with a synthetic heptapeptide, residues GPTIEEVD (HopTPR1-Hsp70$_{GPTIEEVD}$) [306]; 2) PDB ID 3UQ3: ScHopTPR2AB with both MEEVD from the Hsp90 C-terminus, as well as EVD from the Hsp70 C-terminus (HopTPR2AB-Hsp70$_{EVD}$-Hsp90$_{MEEVD}$). In this complex, HopTPR2A interacts with Hsp90$_{MEEVD}$ and HopTPR2B interacts with Hsp70$_{EVD}$ [298]; 3) PDB ID 3UPV: ScHopTPR2B with C-terminal residues PTVEEVD from *S. cerevisiae* Hsp70, SSA4Hsp70 (HopTPR2B-Hsp70$_{PTVEEVD}$) [298].

5.2.2.3 Molecular docking

Protein-protein docking was performed using the High Ambiguity Driven biomolecular DOCKing (HADDOCK) webserver [338]. The best docked structures chosen, based on their HADDOCK score, which is the weighted sum of van der Waals energy, electrostatic energy, desolvation energy, the energy from restraint violations and the buried surface area.

5.2.2.3.1 Hsp70:J proteins

Each J domain model was submitted to HADDOCK along with PfHsp70-x. For each J protein, residues set to be actively involved in binding those forming the HPD motif, and for PfHsp70-x, residues R201, N204, T207, I246 and V419 were set. No passive residues were set for each docking run, but assigned automatically by the server.

5.2.2.3.2 Hop:Hsp90

For these, two sets of docking parameters were used: 1) with active residues set as those that formed interactions in the ScHopTPR2AB-ScHsp90MC template, with passive residues being automatically assigned by the server; and 2) with active residues set as in the most conserved interactions across the modelled complexes and the rest set as passive. The ScHopTPR2AB-ScHsp90MC template was also resubmitted for docking in order to confirm the docking orientation for this complex returned by HADDOCK and also determine the HADDOCK scores returned for this complex orientation.

5.2.3 Identification of important residues in complexes

5.2.3.1.1 Hsp70:J protein

The Protein Interactions Calculator (PIC) webserver [339] was used to identify residue interactions in each protein complex. Default settings were used when submitting models to the PIC server.

5.2.3.1.2 Hsp90:Hop

Complexes for the HopTPR2AB-Hsp90MC interaction (convex analysis) were submitted to the PIC web-server. Python scripts were written to calculate the conserved interacting residues. An interacting residue was considered to be conserved if it interacted in more than half of the complexes observed. Then for each conserved residue, the script would go back to the PIC results and record which residues it interacted with. Using the same conservation

level cut-off, these interactions were used to create conserved interaction networks for the different groupings. Each HopTPR2AB-Hsp90MC complex was also submitted to the Robetta Alanine Scanning web-server [339]. Again, Python scripts were written to record the conserved "hot spot" residues for each grouping. The binding hot spot cut off value used was a $\Delta\Delta G_{bind} \geq 1.0$ kcal.mol^{-1} [340].

5.3 Results and discussion

5.3.1 PfHsp70-x:J domain interactions

Experimental evidence has indicated that PfHsp70-x is exported from malarial parasite, into the host erythrocyte cytoplasm, localising in structures called J dots, along with Type II parasite J proteins, PfA and PfE [260]. The work presented aims to characterise this interaction at a structural level. In addition to PfA and PfE, another Type II J protein predicted to be exported from the parasite, PfB [293], was also included in this analysis. Since PfHsp70-x is exported to the host erythrocyte cytosol, its interaction with potential human J proteins was also considered. These included two type II J proteins, HsDnaJB1 and HsDnaJB4, and two Type III J proteins, HsDnaJC5 and HsDnaJC13; all four of which are expressed within the erythrocyte [341, 342]. Two Type I host J proteins, HsDnaJA1 and HsDnaJA2, were also included in the analysis. These are not expressed within the host erythrocyte, so should not come into contact with PfHsp70-x. As such, they formed a type of negative control, though there is insufficient experimental evidence to suggest these are incapable of interacting with PfHsp70-x.

5.3.1.1 Modelled interactions

The structure of bovine Hsc70 interacting with the J domain of a Type III, J protein, auxillin is available in the PDB, deposited by Jiang *et al*. [291]. To the best of my knowledge, this is the only structure of a J protein interacting with the ATPase domain of an Hsp70 protein (i.e. in a manner that induces ATP hydrolysis). To study this interaction in the context of PfHsp70-x, a hybrid complex modelling approach [335] was used and interactions assessed using the PIC webserver [339]. The results of the nine different J domains interacting with PfHsp70-x, as determined by this hybrid modelling approach are summarised in Figure 5.6. Since this was done for nine different J proteins, the residues are numbered according to the

alignment shown in Figure 5.6B. The interactions displayed were found to be mostly consistent with those reported previously by Jiang *et al*. [291]. From PfHsp70-x, residues I246, F247, I410 and L411 formed hydrophobic interactions with one or more of the J domains modelled. The only consistent hydrophobic interaction here though, was between L411 of PfHsp70-x and P12 from the HPD motif of the J domains, which was found to occur in all nine J domains studied. This residue also formed hydrogen bonds with H11 of the HPD motif, which only occurred in four of the nine modelled complexes, but was the most conserved interaction involving H11. Hydrogen bonds were formed between K29 of the J proteins and residues A418 and V419 of PfHsp70-x, as well as between D13 of the HPD motif and L200 and R201 of PfHsp70-x. There were some interactions found that have not previously been reported, including ionic interactions between D13 of the HPD motif and residues K189 and R201 of PfHsp70-x. Additionally, J protein residue E25 (E26 in Type I human J proteins) was found to interact with K420 of PfHsp70-x, by both ionic interactions and hydrogen bonding. Jiang *et al*. [291] reported an interaction between L200 of PfHsp70-x and H11 of the J domain, which was only seen in the two type III human J proteins. The J protein used by these authors for their structural studies, auxillin, was also a Type III J protein, which may explain this discrepancy. In spite of this, there were few interactions observed that were specific to either type of J protein or in either the PfJ proteins vs their human counterparts. The *P. falciparum* proteins did contain a 3 - 6 residue insertion (residues 16 - 21), when compared to the various human J domains. There were some interactions formed here, obviously absent in the human proteins. The residues here vary greatly and did not form consistent interactions, making unlikely that they play an important role in this interaction with PfHsp70-x, specifically.

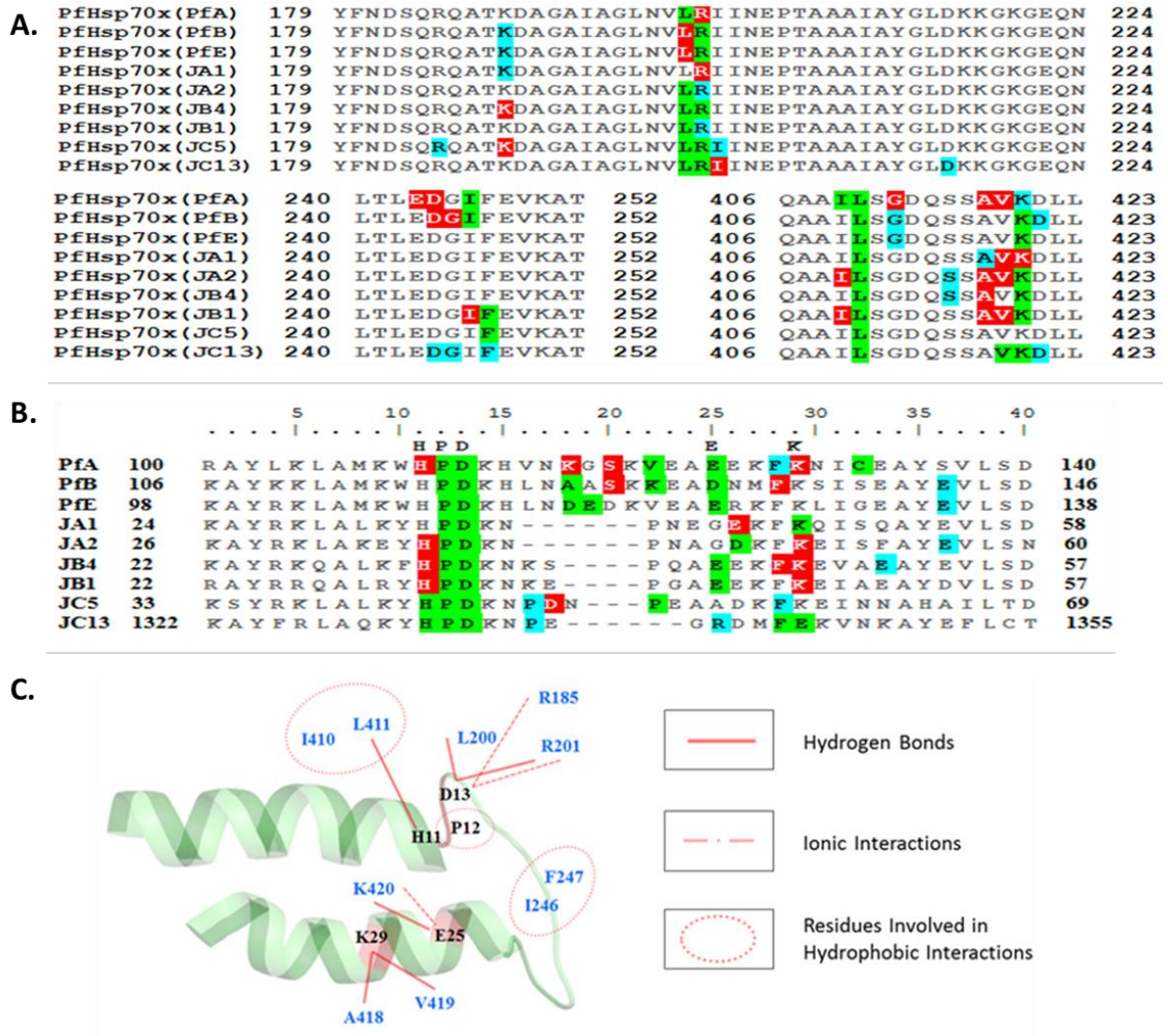


Figure 5.6: Modelled interaction of PfHsp70-x with different J proteins. The aligned sequences of PfHsp70-x (A) and the various J proteins (B) are shown, with residues involved in interactions highlighted as measured in the following complexes. Red: modelled complexes; Blue: minimised complexes; Green: Both modelled and minimised complexes. The human J proteins are labelled as follows. JA1: HsDnaJA1, JA2: HsDnaJA2; JB1: HsDnaJB1; JB4: HsDnaJB4; JC5: HsDnaJC5; JC13: HsDnaJC13. C) Interaction network diagram, representing the conserved interactions which took place between PfHsp70-x and the different J proteins. The portion of the J domain involved in the interaction interface is displayed as a cartoon and residues that form interactions are labelled and coloured black. Interacting PfHsp70-x residues are labelled and coloured blue. Used in Hatherley *et al.* (2014) **[207].**

5.3.1.2 Docked interactions

The Hsp70-J protein interaction structure crystallised by Jiang *et al.* [291] currently presents the only experimentally-determined structure of Hsp70 interacting with the J domain of a J protein. Due to the transient nature of this interaction, the complex could only be determined by cross-linking the residues equivalent to R201 of PfHsp70-x and the aspartic acid of the

HPD motif. This was done because previous work performed by Suh *et al*. [290] indicated that these two residues interact during Hsp70-J domain binding. Jiang *et al*. [291] noted that the J domain could freely rotate around this covalent bond when the structure was solved and it was possible that the orientation of this bound structure may not have been biologically relevant. Additionally, the interactions identified in their structure failed to account of a number of interactions reported previously in literature. Namely residues to R201, N204 and T207 [290], as well as I246 [333] of PfHsp70-x. To further investigate this possibility, each J domain from the modelled complexes was docked to the ATPase domain of PfHsp70-x using HADDOCK [338]. The docking runs were set up to specifically include these residues and the results of this are displayed in Figure 5.7: Docked interaction of PfHsp70-x with different J proteins. All top-scoring complexes displayed favourable interaction energies when scored by HADDOCK. Interestingly, each J domain docked in a different orientation around the PfHsp70-x ATPase domain. Additionally, each of these orientations satisfied the criteria set for docking. A closer look at the various interfaces revealed that each revolved around an interaction between D12 of the J domain HPD motif and that this residue interacted with R201, N204 and T207 of PfHsp70-x. In spite of the various orientations adopted by the different J domains, a number of consistent interacting residues were found in both PfHsp70-x and the different J proteins. Hydrophobic interactions formed with I246 of PfHsp70-x in all docked orientations and most commonly involved residues A7, M8 and F28 of the J domain. Jiang *et al*. [333] performed a number of site directed mutagenesis experiments and found that all tested mutations involving this residue specifically (I216 of BvHsc70) resulted in decreased auxillin binding.

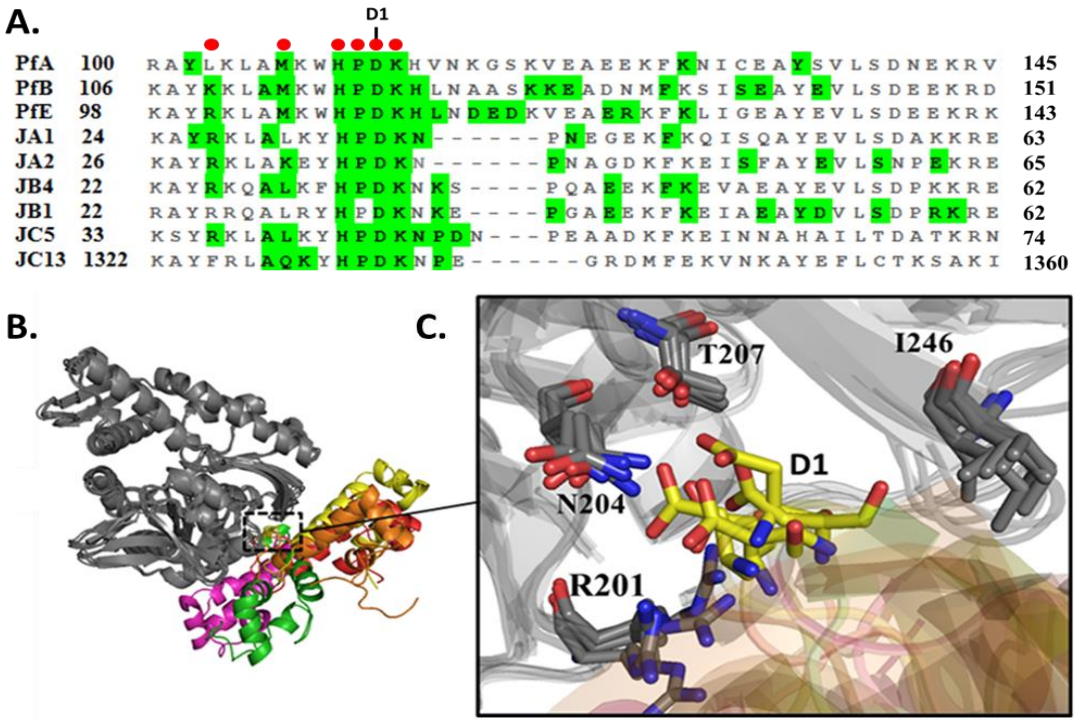


Figure 5.7: Docked interaction of PfHsp70-x with different J proteins. A) Alignment of different J proteins, labelled as in Figure 5.6. Residues involved in interactions with PfHsp70-x are highlighted in green, with the most conserved interacting residues further marked by a red dot. B) A superimposition of the ATPase domain of PfHsp70-x in the different complexes. A zoomed in view of the region indicated by a dotted border is shown in (C). Residues of PfHsp70-x are shown as sticks, coloured grey and residue D1 of each J protein is also shown as a stick and coloured yellow. Adapted from Hatherley *et al.* (2014) **[207].**

The results of computational alanine scanning (Table 5.1) identified I246 as a hot spot residue ($\Delta\Delta G_{bind} \geq 1.0$ kcal.mol^{-1}), with the exception of the interaction with PfE, which returned a $\Delta\Delta G_{bind}$ of 0.94 kcal.mol^{-1} for this residue. The most consistent hydrogen bonds involved R201, N204 and T207 of PfHsp70-x. As mentioned before, each of these interacts with D13 of the HPD motif by forming hydrogen bonds and R201 forms ionic interactions D13. Again the exception to this was the PfE complex, in which R201 did not interact in any way with D13. Only R201 of these three PfHsp70-x residues was identified as a hot spot residue across all complexes. The complexes with HsDnaJA2 and HsDnaJB1 identified R201, N204 and T207 as hotspot residues. Two adjacent acidic residues of PfHsp70-x, E243

and D244, formed both hydrogen bonds and ionic interactions with either R4 or K14 of the J proteins.

Table 5.1: Alanine Scanning of PfHsp70-x docked with different J proteins. Results are shown for PfHsp70-x (A) and J protein residues (B), as measured in complex. J proteins are named as in Figure 5.6 and values indicating hot spot residues are coloured red. Adapted from Hatherley *et al.* (2014) **[207].**

A.

70-x	PfA	PfB	PfE	JA1	JA2	JB4	JB1	JC5	JC13
R185	-0.14	-	2.09	-	-	-	-	-	-
R201	2.46	2.70	3.18	2.69	1.49	2.10	2.53	2.49	2.04
N204	0.55	-0.09	3.18	0.86	1.75	0.82	3.94	0.93	0.95
T207	0.69	0.58	0.82	0.65	1.60	0.73	1.46	-0.04	0.64
D216	-	-	-	-	-	0.89	0.14	0.09	0.08
E243	-0.16	0.71	-	-	0.15	-0.02	0.25	0.56	-0.02
D244	-0.12	-	-	0.69	-	0.96	-	3.13	-0.08
I246	1.62	1.41	0.94	1.07	1.38	1.34	1.55	1.96	1.58
F247	0.97	0.00	0.33	0.29	0.27	1.13	0.41	0.94	1.13
E248	-	1.70	0.82	-	2.60	-	-0.03	0.01	-
Q407	0.17	0.60	0.17	2.62	-	1.00	-	0.64	0.53
L411	0.33	-	0.40	1.13	-	1.07	0.30	0.65	1.17
V419	0.35	-	-	0.11	-	-	0.58	-	0.49
K420	2.44	-	-	1.64	-	1.13	1.06	0.65	1.37
D421	-	-	-	-0.11	-	-0.08	-0.07	0.49	-

B.

Alignment Position	PfA	PfB	PfE	JA1	JA2	JB4	JB1	JC5	JC13
3	1.74	1.15	-	0.28	-	-	-	-	-
4	0.64	0.74	-	1.14	1.78	1.27	-	5.43	-
8	0.87	-	0.18	0.80	1.86	1.18	-	1.39	1.60
11	1.78	1.54	0.62	2.11	0.74	1.65	1.13	2.25	2.26
13	2.71	1.90	1.98	2.77	3.48	2.00	0.85	1.46	2.42
14	0.18	0.13	1.80	2.34	1.40	1.97	1.28	2.41	0.78
15	-	-	-	0.93	-	-	-	-	-
16	-	-	0.79	-	-	1.93	1.05	-	-
18	-	-	3.41	-	-	-	-	-	-
25	-	-	2.23	-	-	0.59	-	-	-
28	0.79	1.05	-	0.65	-	-	-	-	-
29	-	-	1.44	-	-	-	1.05	-	-
32	-	2.16	-	-	-	-	-	-	-
35	0.91	-	-	-	1.18	-	-	-	-
36	-	-	-	-	3.64	-	-	-	-
40	-	-	-	-	1.78	-	-	-	-
42	-	-	-	-	1.44	-	1.84	-	-

Much like the HPD motif, R4 of the J domain is conserved as a positively charged residue (as R or K) and its importance has been reported in various J proteins [292, 343–345]. Residue

K14 took part in a number of interactions with additional residues, including D216, D248, K420 and D421. This is understandable, since it is adjacent to D13, around which the difference orientation formed and in most complexes this residue was identified as a hot spot residue. Of the different residues K14 interacted with, either E248 or K420 were identified as hotspot residues in most complexes. Residue H11 of the J domain took part in a number of varied hydrogen bonds across the different docked complexes and was identified as a hot spot residue in all complexes in which it formed an interaction.

5.3.1.3 Comparison of modelled complexes to docked complexes

The initial approach undertaken in this work involved using hybrid modelling to study the interaction interface of PfHsp70-x with potential J proteins, based on the BvHSc70:auxillin template structure solved by Jiang *et al*. [291]. This was supplemented with molecular docking using HADDDOCK, guided by residues highlighted in literature as being important to the Hsp70-J protein interaction. The docking results yielded a substantially different set of interactions to those found by complex modelling. On inspection, the HPD motif formed interactions in a similar area in both interaction sets, but residue D13 was buried deeper within the ATPase domain interface. This enabled interactions to be formed with N204 and T207. In their study, Jiang *et al*. [291] reported a reduced binding affinity between BvHsc70 and the auxillin J domain, as prepared by disulfide-linking (refer to Section 5.2.2.2.2). It is possible that this cross-linked interaction prevented the J domain from completely entering the binding groove in the ATPase domain. There are a number of features that support the interactions described by docking over those obtained through hybrid modelling. Firstly, since these were specified in the docking parameters, the active involvement of I246, N204 and T207 in this set of interactions, means that these agreed more closely with work performed by Suh *et al*. [290] and Jiang *et al*. [333]. Since the HPD motif comprised the only active residues specified when docking, it was interesting to find a number of other residues

in the interaction interface that, through review of literature, have experimental evidence which suggests their importance in the Hsp70-J protein interaction. These include R4 [292, 343–345], K14 and F28 [292]. Another interesting find was an NMR-based structural study, performed by Ahmad *et al*. [346]. This focused on J domain interactions with Hsp70 in its ADP-bound state (i.e. after stimulation of ATP hydrolysis). Their results indicated that in the ADP-bound form, the interaction interface shifts from the HPD motif to the positively-charged residues of Helix II of the J domain, which interact with acidic residues of the loop of Hsp70. These acidic residues of this loop surround the residue corresponding to I246 of PfHsp70-x. Helix II of the J domain used in the present study comprises residues R1 – M8. Results presented here indicate that A7 and M8 form hydrophobic interactions with I246 of PfHsp70-x and that R4 forms both hydrogen bonds and ionic interactions with E243 and D244. Therefore, this more buried complex interface includes interactions involving more residues previously found to be important in Hsp70-J protein binding and also describes a pose in which The J domain interface can switch to interact with the acidic loop of Hsp70.

5.3.1.4 Hsp70 conformation during J protein docking

The two templates used to model the full-length structure of PfHsp70-x were 1YUW and 2KHO, which represent this protein in its ATP- and ADP- bound state, respectively. The nine different docked orientations were superimposed to these two models to provide further evidence to suggest which the correct binding orientation is. All docked orientations clashed with the SBD of PfHsp70-x in its ATP-bound state (Figure 5.8A). In this structural conformation lid of the SBD interacts with the residues surrounding I246 of PfHsp70-x [333]. Taken together with the finding that J domain stimulation regulates the inter-domain communication between the SBD and the ATPase domain [291], this would suggest that the J domain competes with the SDB of Pfhsp70-x to interact with the ATPase domain by binding to these residues and displacing the lid of the SBD. The findings of Ahmad *et al*. [346], that

the J domain interaction shifts to focus mainly this site in the ADP-bound state of Hsp70 further support this.

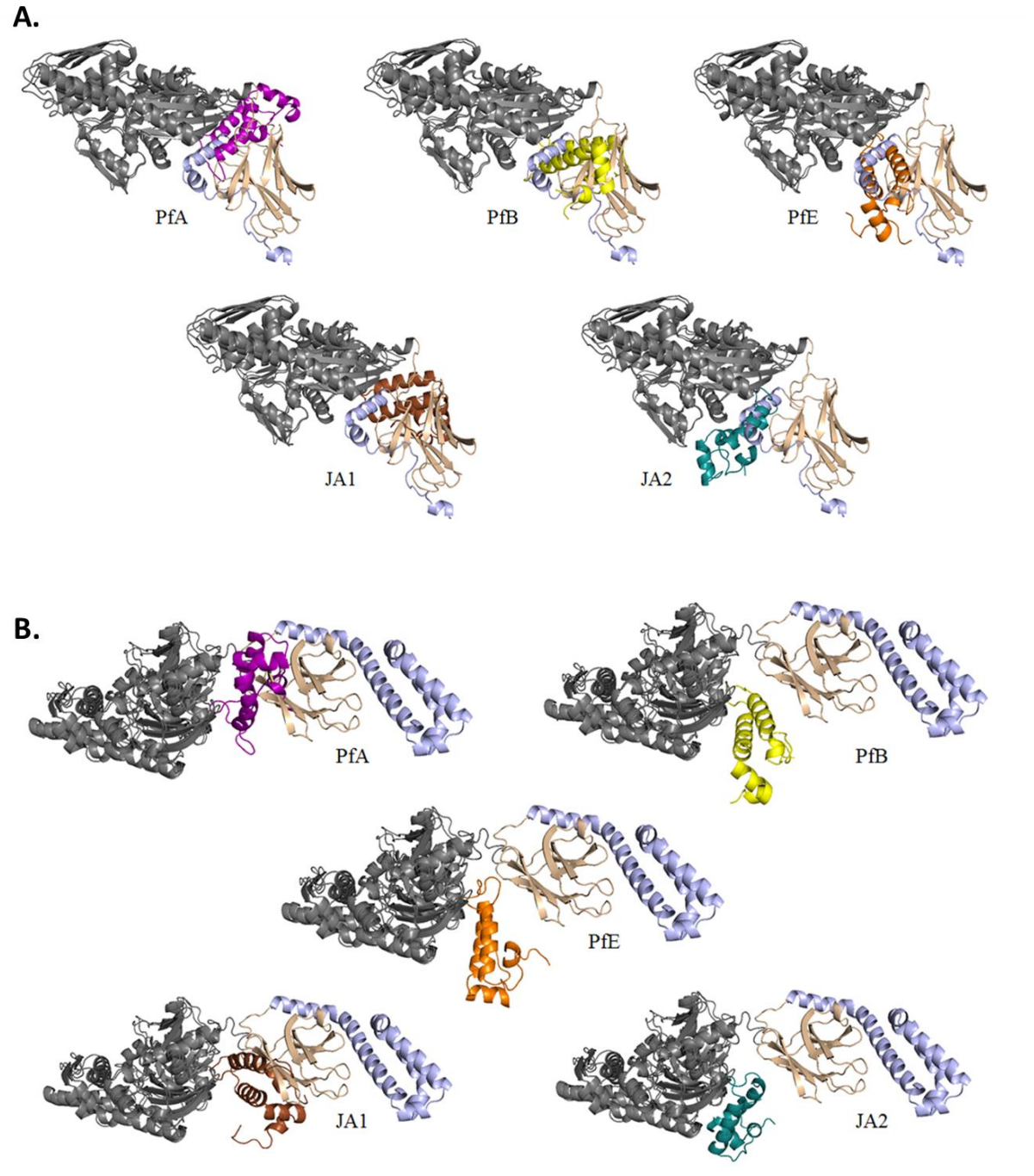


Figure 5.8: Docking orientations superimposed to PfHsp70 in both ATP and ADP-bound states. The docked complexes for the three PfJ proteins, as well as two host J proteins, in complex with PfHsp70-x are shown, superimposed with the full-length structure of PfHsp70-x, modelled in its ATP-bound state (A) and ADP-bound state (B). J domains are coloured as follows: PfA - purple; PfB - yellow; PfE – orange; JA1 - brown and JA2 - teal. Used in Hatherley *et al.* (2014) **[207].**

In the ADP-bound state of PfHsp70-x, only complexes with PfB, PfE and HsDnaJA2 did not clash with the SBD (Figure 5.8B). It is tempting to argue that one of these represent the correct docked orientation. The interactions of the J domain with the ATPase domain of Hsp70 are, however, highly dynamic [346] and the shift away from interactions with the HPD motif and toward helix II of the J domain may change this.

5.3.1.5 Correct orientation for docking

The presence of different docking orientations makes it difficult to characterise this interaction. Unfortunately each docked orientation had its merits. Only complexes with HsDnaJA2 and HsDnaJB1 had all four residues specified in the docking, R201, N204, T207 and I246, as hot spot residues. Complexes with HsDnaJA1, HsDnaJA2, HsDnaJB4 and HsDnaJC4 identified residue R4 of the J domain as a hot spot residue and complexes with PfA and PfB had Y3 as a hot spot residue, which is also supported by Genevaux *et al*. [292]. Superimposition to the ADP-bound state of PfHsp70-x, suggest PfB, PfE and HsDnaJA2 as promising orientations. Again, the dynamic nature of this interaction may suggest a number of these interactions may be correct. HsDnaJA2 does seem to be the most promising though, as these interactions include the most residues supported by literature and it did not clash with the SBD of PfHsp70-x when superimposed to this protein in its the ADP-bound state.

5.3.2 Hsp90:Hop interactions

The second set of interactions studied were those between Hsp90 and Hop. Based on structural work performed by [298], a comparative study was undertaken to characterise the Hsp90:Hop convex interaction interface between both human and *P. falciparum* proteins. Further, the better characterised concave interaction interface was also investigated and compared between human and *P. falciparum*.

5.3.2.1 Convex interaction interface

The convex interaction interface template consists of both the M and C domains of Hsp90 and Hop TPR2A and TPR2B domains. This complex will be referred to as Hsp90MC-HopTPR2AB. As with the PfHsp70-x-J protein interaction analysis, Hsp90MC-HopTPR2AB complexes were generated using both a modelling and docking approach and done for PfHsp90 (PfHsp90MC-PfHopTPR2AB), HsHsp90-alpha (HsHsp90αMC-HsHopTPR2AB) and HsHsp90-beta (HsHsp90βMC-HsHopTPR2AB), with their respective potential Hop counterparts. Unlike the Hsp70:J-protein complexes, both the docked and modelled complexes of the Hsp90:Hop interaction interface formed in the same orientation. Since there were four different complex models representing each Hsp90:Hop interaction interface (two from modelling and two from docking), the results of the PIC web-server protein-protein interactions were analysed using python scripts. The results were summarised to create an interaction network diagram for each interaction interface (Figure 5.9).

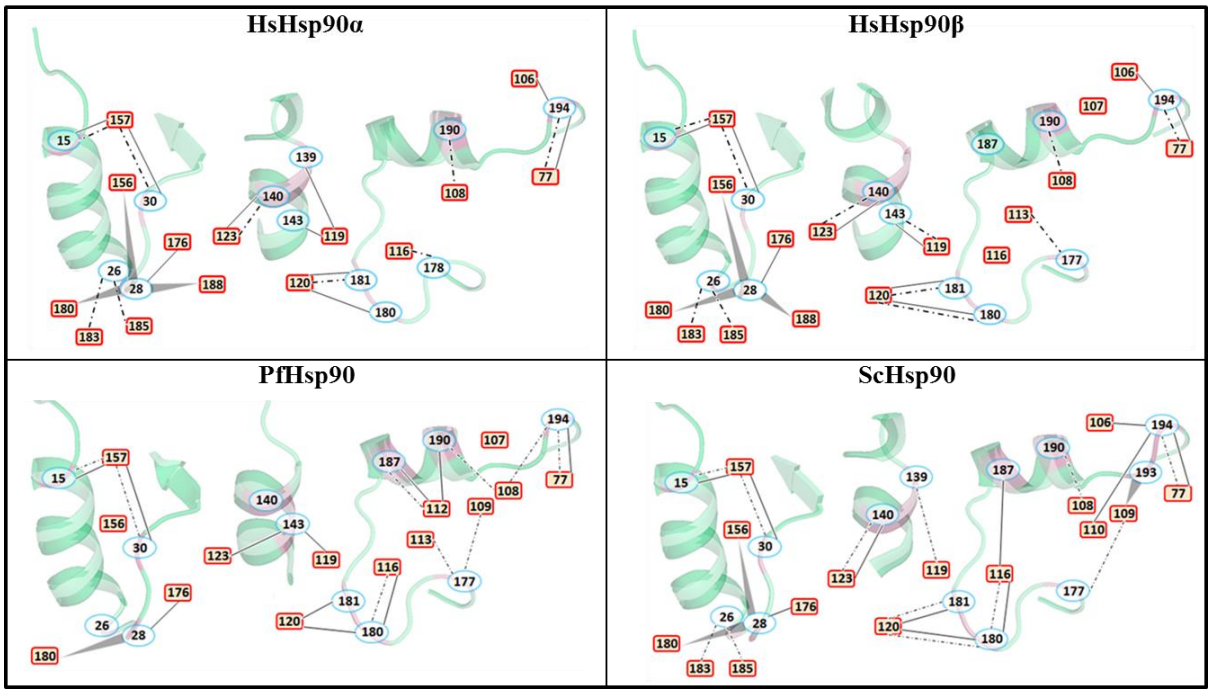


Figure 5.9: Hop:Hsp90 convex interaction interface. An interaction network is shown for the Hop:Hsp90 convex interaction interface. The Hsp90 M domain is shown in cartoon, with interacting residues shown as blue ovals. Hop residues are indicated as red rectangles. Interactions are shown as follows. Solid line: ionic interaction; Dotted line: hydrogen bonding; Triangle: hydrophobic interactions. Used in Hatherley *et al.* (2015) [208]**.**

Although both the M and C domains of Hsp90 were present in these models, the interaction interface was limited to Hsp90 M domain only. The common interactions between HopTPR2AB and the Hsp90 M domains across the different sets of complexes were quantified and presented in Table 5.2. A more detailed list of these interactions is available in Appendix G; Table G1 and Table G2.

Table 5.2: Total interactions common and unique to convex Hsp90:Hop interaction interfaces. The number of common interactions observed was totalled and summarised. A) Comparison of human complexes to the yeast template used in this study. B) Comparison of the human complexes with *P. falciparum* complex. The conserved column indicates that the interactions were common to all three complexes. Complexes are named, based on the Hsp90, as follows. Alpha: HsHsp90α; Beta: HsHsp90β; Pf: PfHsp90; Template: ScHsp90. Adapted from Hatherley *et al.* (2015) [208]**.**

A) Human complexes vs yeast template complex					
Conserved	**Alpha & Beta**	**Template & Beta**	**Template only**	**Alpha only**	**Beta only**
18	3	1	7	1	4
B) Human complexes vs *P. falciparum* complex					
Conserved	**Alpha & Beta**	**Pf & Beta**	**Pf only**	**Alpha only**	**Beta only**
12	9	2	11	1	3

When comparing the two human complexes to the yeast template, the majority of interactions were common to all three complexes and seven interactions were found only in the yeast template complex (Table 5.2A). Thereafter, few differences were found when comparing these three complexes. When considering the PfHsp90 complex and the two human complexes, the number of interactions common to all three interfaces drops to 12 (Table 5.2B). Nine interactions found in both human Hsp90 complexes were not present in the PfHsp90 complexes and 11 interactions were specific to the PfHsp90:Hop interaction interface.

5.3.2.1.1 Common interactions in the convex interface

Again, to make the different complexes comparable, the residues were numbered according to their positions in an alignment between the different sequences. The concave interaction

interface of Hsp90 and Hop is not very well characterised and a structure describing this interaction was only published in 2012 [298]. The complexes modelled and docked in the present work provide an interaction interface for both *P. falciparum* and human proteins. Combining the interaction data of these with the yeast template a general interaction interface has been predicted. A number of interactions were common to all four sets of complexes (yeast, human α-, β- and Pfhsp90s with their respective Hop counterparts). Primarily, for all complexes, ionic interactions occurred between Hsp90-E15 and Hop-K157, Hsp90-D30 and Hop-K157, Hsp90-D180 and Hop-R120 (K120 in HsHop and ScHop), Hsp90-E181 and Hop-R120, as well as Hsp90-E194 and Hop-K77 (for actual residue numbers, refer Appendix G; Table G3 and Table G4). Many of these also formed hydrogen bonds. Cation-π interactions also occurred between Hsp90-W28 and Hop-R176. Residue W28 (W300 in ScHsp90) has been previously highlighted in literature as being important to the interaction with Hop in yeast [347]. Hsp90-W28 also formed a hydrophobic interactions with L180 of Hop (Y180 in HsHop). The alanine scanning results identified this residue as a hot spot residue in the human and yeast complexes, but not in the *P. falciparum* complexes (this difference is discussed in 5.3.2.1.2). Interestingly, there were also conserved hydrogen-bond interactions, despite a lack on conservation at a residue level across the different species. Residue D190 of PfHsp90 (present as S108 or T108 in the other HsHsp90s and ScHsp90), formed hydrogen bonds with N108 of PfHop (present as T108 in both HsHop and ScHop). This was the only exception where such a lack of conservation between the human, yeast and *P. falciparum* proteins did not cause a difference in the way they interacted. There was just one interaction common to the human and *P. falciparum* complexes, but not seen in the yeast template. This was an ionic interaction between K143 of Hsp90 and E119 of Hop.

5.3.2.1.2 Differences between *P. falciparum* and human complexes

As mentioned above, the human Hsp90:Hop complexes displayed interactions not seen in their *P. falciparum* counterparts and vice versa. Some of these had no obvious cause, as with the hydrogen bonding between Hsp90-N26 with both L183 and E185 of Hop. All of these residues are conserved in both the human and *P. falciparum* systems. There were two hydrophobic interactions with Hsp90-W28 only seen in the human complexes, involving residues M156 and L188 of HsHop. In PfHop these are polar residues K156 and S188 and so did not form these hydrophobic interactions with PfHsp90-W28. The absence of these four interactions involving N26 and W28 of PfHsp90 may explain why W28 was not identified as a hot spot residue through alanine scanning. It is possible that the W28 interactions are not as crucial to the stability of the PfHsp90:PfHop complex as with the human and yeast complexes. The importance of this residue has so far only been indicated in yeast [274, 347] and it would be interesting to test its importance in both PfHsp90 and the two HsHsp90s. Another noted difference was the hydrogen bonding and ionic interactions between HsHsp90-E140 and HsHop-K123. The latter of these residues is present as PfHop-E123, which instead forms an ionic interaction with PfHsp90-K143. The final interaction conserved in the human complexes was an ionic interaction between HsHsp9-E194 and HsHop-H106. Again the residue is different in PfHop, present as D106 and formed no interactions with PfHsp90. The interactions seen in the *P. falciparum* complexes and not their human counterparts were more blatant and could be attributed to residue differences in either PfHsp90, PfHop or both. These included a hydrogen bond between PfHsp90-K177 and PfHop-E199. This was due to an unconserved Hsp90 residue, present as A177 in HsHsp90α and Q177 in HsHsp90β. PfHsp90-D180 formed both hydrogen bonds and ionic interactions with PfHop-K116 (Q116 in HsHop). Ionic and hydrogen bonds also occurred between PfHop-R112 (L112 in HsHop) and both PfHsp90-E187 and PfHsp90-D190 (T190 and S190

in HsHsp90α and HsHsp90β, respectively). Finally, PfHsp90-E194 formed hydrogen bonds with PfHop-N108 (T108 in HsHop). These final two sets of interactions may be the most significant in this study. Residues PfHop-N108, PfHop-R112 and PfHsp90-E194 were identified as hot spots, which was not found to be the case in their human counterparts (Table 5.3). These may be worth testing experimentally as they could present good target sites for drug development.

Table 5.3: Hot spot residues from the Hsp90:Hop convex interaction interface. Hot spot residues are indicated for the different complexes. Residues are numbered as in the models used to characterise these interactions, with the actual residue numbers shown in brackets. Complexes are named as in Table 5.2. Used in Hatherley *et al.* (2015) [208]**.**

Hsp90 Residues			
Template	**Alpha**	**Beta**	**Pf**
22 (K249)			
28 (W300)	28 (W320)	28 (W312)	
178 (S450)	178 (S470)		178 (S489)
180 (D452)			
191 (R463)	191 (R483)	191 (R475)	191 (R502)
194 (E466)			194 (E505)
Hop Residues			
			108 (N350)
			112 (R354)
116 (R376)	116 (Q340)		
	160 (T284)	160 (T284)	
176 (R436)	176 (R400)	176 (R400)	176 (R418)

5.3.2.2 Concave interaction interface

The concave sites of Hop are interact with the C-terminal peptide regions of both Hsp70 and Hsp90 [298, 299, 304–308] and is the target site for drugs that target Hop:Hsp70 or Hop:Hsp90 interaction [314, 318, 321]. These drugs have been developed as part of cancer treatments and therefore target human proteins. The work presented here involves identifying differences between human and *P. falciparum* chaperone systems as potential targets for drug

development. As such, the concave interface has also been investigated in order to determine how well suited it is to this purpose. The results of this analysis indicate a number of interactions in each complex that in each common to both human and *P. falciparum* (Table 5.4). On the other hand, relatively few were specific to only human or *P. falciparum*.

Table 5.4: Total interactions common and unique to concave Hsp90:Hop interaction interfaces. The number of common interactions observed was totalled and summarised for the different concave interaction interfaces studied. Adapted from Hatherley *et al.* (2015) [208]**.**

TPR1-Hsp70$_{GPTIEEVD}$		
Human and Pf	**Human only**	**Pf only**
15	8	2
TPR2A-Hsp90$_{MEEVD}$		
Human and Pf	**Human only**	**Pf only**
13	4	5
TPR2B-Hsp70$_{EVD}$		
Human and Pf	**Human only**	**Pf only**
6	1	0
TPR2B-Hsp70$_{PTVEEVD}$		
Human and Pf	**Human only**	**Pf only**
11	3	1

An even more interesting result was obtained when comparing results across the different TPR domains (Figure 5.10). There were several interactions conserved even across these different TPR domains (TPR1, TPR2A and TPR2B). This indicates that the manner in which Hop interacts with the C-terminal peptides of Hsp70 and Hsp90 is highly conserved across different species and different TPR domains. This explains why it is a promising target for anticancer drug development. The potential of this interaction interface as a target site for selective inhibitors of the PfHsp90:PfHop or PfHsp70:PfHop complex interaction may be more limited.

Figure 5.10: Alignment of different TPR domains, highlighting common Hsp70/Hsp90 C-terminal peptide contact points. Residues from each TPR domain which formed interactions in the complexes studied are highlighted in green. Used in Hatherley *et al.* (2015) [208]**.**

5.4 Conclusion

Through the analysis of interaction between PfHsp70-x and various J domains, the docking results appeared to account for a more complete interaction interface than the modelled complexes. This interface contained interactions involving residues that were supported by experimental data and fits the data suggesting a shift toward helix II of the J domain after J protein-stimulated ATP hydrolysis. The main problem faced was identifying the correct docking orientation from the different HADDOCK results and there was insufficient data to make this decision. In spite of this, a number of residues were found to consistently be involved in this interaction and, once again, many were supported by literature. Part of this study was aimed at finding ways to target the PfHsp70-J protein interaction. Unfortunately all residues identified to be important in this interaction occur in highly conserved regions of the protein [207]. Therefore targeting the ATPase domain of PfHsp70-x does not appear to be the correct interface for targeting this interaction. Both Botha *et al*. [296] and Hennessy *et al*.

[344] highlight that J domain interactions display promiscuous selectivity, even with Hsp70s from different organisms. It may be interesting to expand on this study by including both human Hsp70s and the other cytosolic *P. falciparum* Hsp70, PfHsp70-1. This may help to better characterise the Hsp70-J domain interaction interface for comparative purposes. The high level of conservation at the different J domain contact points suggests that these would not provide any new information. However, including additional analysis techniques, such as molecular dynamics or Perturbation-Response Scanning [348], may reveal features of this interaction that was missed by molecular docking.

The results from the Hsp90:Hop analyses were more promising. When compared to the concave interaction sites, the interaction interface at the convex sites were far less conserved between human and *P. falciparum* proteins. In addition to this, a number of hot spot residues were identified specifically in the malarial complexes and not in their human counterparts. These residues may be potential site for drug development, aimed at specifically inhibiting the Hop:Hsp90 interaction of *P. falciparum*.

Chapter 6 Automated homology modelling tool

Homology modelling is a fairly well-established technique that has become a fundamental tool in bioinformatics for studying proteins which have no solved structures available in the PDB. A number of automated protein modelling servers are available, but each has their own limitations. Presented in this chapter is a simple set of scripts that automate various aspects of the homology modelling process, while still allowing user input where desired. A simple benchmarking test was run to assess the reliability of the tool and accuracy of models produced using the tool. A web interface is also currently under development to provide academic users with access to this tool for their own research purposes. This is presented to show how the modelling tool can be integrated into a web-based application that allows users to intuitively and reliably model target proteins of interest.

6.1 Introduction

Homology modelling provides a means to obtain structural information about a protein, where this has not been determined experimentally, by means of X-ray crystallography, NMR or EM) [185]. As with other theoretical approaches to protein structure determination, homology modelling works primarily on the concept established by Epstein and co-workers in 1963 [349], that the structure of a protein is determined by its amino acid sequence [185]. This technique also relies on the observation that the structure of a protein is far more conserved than its amino acid sequence. While this holds true the longer a protein sequence is [183], it has been shown that the structure of a protein can be up to ten times more conserved than its sequence [182].

6.1.1 Steps involved in homology modelling

The process of homology modelling can be divided into seven different steps: 1) Template identification; 2) Sequence alignment; 3) backbone modelling; 4) loop modelling; 5) side-chain modelling; 6) structural refinement and 7) model evaluation [184, 185].

6.1.1.1 Template identification

The selection of a template is greatly important step in the homology modelling process [184, 350]. The Protein Data Bank (PDB) is an online resource containing the structures of proteins and nucleic acids that have been solved experimentally [351]. The database currently (June 2015) contains 101,356 protein structures and is updated on a weekly basis [352]. These, however, account for only 35,747 unique protein sequences. The challenge of template identification is to find a protein with known structure that is similar to the target protein of interest [350]. Initially, only the sequence identity between target and template was used when making this decision [183–185]. Some empirical guidelines were established to help with this, the most well-known being that proteins that share more than 40% sequence identity will have similar structures, which is true with few exceptions [184]. Below this "safe zone" [185], is the so-called "twilight zone", which describes another threshold at which it is unclear whether or not two proteins are homologous [183]. This is usually set between 20 – 35% sequence identity, depending on the length of the alignment between the two sequences. The Basic Local Alignment Search Tool (BLAST) [353] is commonly used to identify templates [185, 350], especially if only sequence information is considered. This uses substitution matrices to score alignments which positively scores conserved substitutions and penalises gaps and poorly substituted residues when comparing sequences. BLAST assesses local similarity using a word-based approach, which aligns short segments (default length 3 residues for protein BLAST) from the query sequence against sequences in a database, and scoring these short alignments. If a segment pair scores above a certain threshold, it is extended and rescored, until it falls below the threshold. This approach aims to identify a maximal scoring pair (MSP), which is the highest scoring aligned region between to sequences of the same length. This way it is able to quickly search the large number of protein sequences and return those which are most similar to the sequence queried. In 1997, a Position-Specific Iterated BLAST (PSI-BLAST) was released which uses the output from

BLAST to construct a position-specific score matrix (PSSM), which it uses in subsequent searches to find more distantly related homologues [354].

At best, this approach will provide a set of potential templates, along with their sequence identity to the target. The use of a substitution matrix does provide more information about the evolutionary relationship between two sequences, rather than just looking at sequence identity; however, improved template identification can be achieved by incorporating structural information into a template search, such as secondary structure prediction [184]. Secondary structure prediction simply distinguishes whether regions of a protein sequence form an α-helix, β-sheet or disorder regions [355]. Another successful method incorporates the use of Hidden Markov Models (HMMs) [356, 357], as with the HHPred server [326]. This program builds profile HMMs using the results from a PSI-BLAST search, as well as secondary structure information either predicted using PSIPRED [358] or assigned from PDB structures using DSSP [359]. This contains position specific information on each residue in a sequence, with an assigned probability for each possible amino acid substitution, as well as insertions, deletions and gap extensions. The HMM is then compared to a precompiled set of HMMs for each structure searched against when identifying potential templates for modelling [326].

Template evaluation is another consideration when identifying a suitable template for modelling [184, 350]. It is important to look at the quality of the structure solved. The PDB currently produces validation reports for its X-ray crystallographic entries [360], which are summarised when an entry is viewed. For X-ray crystal structures, the resolution can be a good, quick measure of the structure's quality when assessing a potential template [350]. Factors relating to biological relevance may also come into consideration, such as the presence or absence of bound ligands or substrates. It can be helpful comparing the predicted secondary structure of the target to the assigned and predicted secondary structure of the

template, as provided by HHPred [326]. Finally, query coverage (how much of the target sequence is accounted for by the 3D structure of the template) is important to look at since some structures have only been partially solved. This information is returned when searching for templates using programs such as BLAST and HHPred.

6.1.1.2 Sequence alignment

Once a template is obtained, amino acids from the target sequence need to be mapped to the 3D structure of the template (described in Section 6.1.1.3), which is done based on the alignment between the target and template sequences [184, 328, 361–363]. Westhead & Thornton [364] reviewed the original CASP results (CASP1 and CASP2) and found that a correct sequence alignment most greatly affected the accuracy of models for comparative modelling. Reports for both CASP4 and CASP7 highlighted alignment accuracy as an important hurdle that needed to be overcome in order to improve the quality of predictive models [365, 366], though there has been steady progress in successive CASP experiments [367].

Many different sequence alignment programs have been developed, each with different algorithms to improve alignment accuracy (Section 6.1.2). There are, however, a number of different techniques that can be applied to ensure a better alignment. In CASP3 the use of multiple sequence alignments (MSAs), as well as the use of structural data was attributed to greatly improve alignment accuracy, when compared to CASP1 and CASP2 [368, 369]. In the case of low sequence identity, MSAs help to identify positions in the sequence where certain types of residues are conserved through evolutionary processes, which may not be apparent when simply aligning two sequences [185, 370]. Likewise, consulting the 3D structure of the template and other homologous proteins can help to positions gaps in the alignment such that they are easier to model [185]. Some alignment algorithms, such as

Promals3D [327] and 3D Coffee [371, 372], are designed to take structural information into account when performing sequence alignments.

6.1.1.3 Modelling backbone structure

This involves fitting the amino acid sequence of the target to the 3D structure of the template [184, 328, 361–363]. The process starts with assigning the backbone of the protein structure, followed by improving the loop regions and positioning side chains of each residue. There are a number of different methods to predict the backbone of the homology model. The original method of comparative modelling was the assembly of rigid bodies [373]. This involves superimposition of a number of homologous structures and taking main chain positions as an average value for that position in the different structures. It still forms part of a number of homology modelling algorithms [374–376]. Another method, segment matching, involves assembling small fragments of protein structure together to build a protein piece by piece [373, 377]. Satisfaction of spatial restraints is the method used by MODELLER [328]. The spatial restraints are presented in the form of a probability density function (PDF). The PDF is used to predict the most probable 3D structure of a protein, based on Phi/Psi angles, Cα-Cα distances, etc., taking residue type into consideration. The protein is modelled in such that the molecular PDF is optimised [328].

6.1.1.4 Loop refinement

Loops are considered to be parts of a protein that do not have a rigid secondary structure, such as alpha helices or beta strands [164]. They still have been found to play important biological roles. The antigen-binding sites of antibodies contain antigen binding loops that may confer specificity [378]. Serine proteases cleave specific loops in target substrates, which are mimicked as a part of certain inhibitor design strategies [379]. Many proteins also contain loop regions responsible for binding metal ions [380]. Additionally, if sections of a model cannot be determined from the template structure, due to insertions or deletions, these

are modelled as loops with structures determined using *ab initio* methods [381]. These are determined using database approaches or search-based methods to predict the structure [164]. These use loops from known structures in the PDB to predict the loop in question, since the PDB structure's loop represents that amino acid sequence in its 'natural' conformation. This method is considered to be more limited than others when predicting loops, though the limit of length of loops predicted using this method is 15 amino acids [164, 382]. Search-based methods use conformational sampling and energy calculations to predict the structure of a loop, based on the knowledge of the physicochemical properties of amino acids. These approaches are often supplemented with database search methods [164]. MODELLER provides a loopmodel class for loop refinement [383]. This involves optimising a statistical energy function developed specifically for the prediction of accurate loop conformations [384].

6.1.1.5 Side chain positioning

Determining the correct positions of amino acid side chains within a modelled 3D protein structure is an incredibly difficult and computationally expensive task [385]. This process usually involves the use of rotamer libraries [386], which contained preferred side-chain conformations for each amino acid [387]. The use of this library meant that fewer conformations had to be searched, making the process far less computationally expensive [387]. Still, this requires the best rotamer to be calculated for each residue in the protein [184]. Due to this, heuristic methods are used, since they are relatively fast, though they will return an answer that is the local minimum and not necessarily the most correct [388]. With MODELLER, side chain positioning is calculated as part of the spatial restraints, including all bond angles, as well as preferred distances from the main chain atoms and other side chain atoms [383].

6.1.1.6 Structural refinement

This is an area of modelling that still requires a great deal of progress to be made [184]. Refinement of protein structures use an approach proposed by Levitt & Lifson [389], which involves refining the structure of a protein model, using a generalised force field, until it is in its most energetically favourable conformation. By default, MODELLER performs structural refinement as part of the modelling process [383]. After optimising the objective function, models are refined using molecular dynamics with simulated annealing [328].

6.1.1.7 Model quality assessment

In the absence of an experimental structure it is impossible to tell how closely a modelled protein resembles its native state [366]. This is why in the CASP experiments there is a great deal of focus on model quality assessment programs. Being able to distinguish between a good and a bad quality model is an aspect that is key to developing a strong protein modelling method. Model quality assessment programs are usually developed by either computing physicochemical properties or by observing the differences between theoretical models and experimental structures [390]. As with all aspects of modelling, there are many different model quality assessment programs (MQAPs) available (Section 6.1.3). Each approach has its own strengths and limitations, so when evaluating models a number of different programs should be used [329].

6.1.1.8 Repeat where necessary

After evaluating the model, it may not display sufficient quality and will need to be modelled again, by changing certain parts of the modelling process [350]. Every step in the homology modelling process will influence the overall quality of the protein model, although the earliest steps will have the greatest effect [391]. Most often, improving model quality can be achieved by adjusting the alignment [350]. However, choosing to use different modelling software, as well as performing structural modifications can also have positive results [374].

The example of structural modification provided by Wallner *et al* [374] is use of the software SCWRL [392], which can be used to improve side chain positioning after modelling has been completed.

6.1.2 Sequence alignment algorithms

Protein and nucleic acid sequence alignments are a central to the field of bioinformatics [370]. Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment).

6.1.2.1 Clustal

There are a number of different MSA programs released in the Clustal series and are widely used for global MSA [370]. These have undergone a number of changes to improve the quality of alignment produced. Clustal uses a progressive alignment strategy. This involves comparing two sequences at a time to progressively build the MSA. The order in which these alignments are performed is based on a guide tree, created using phylogenetic-based algorithms. In order to improve the algorithms used by Clustal, up-weighting of more divergent sequences was implemented to reduce the biases caused by near-identical sequences [393]. Position-specific gap penalties attempt to introduce gaps in loop regions rather than portions of the sequence that form secondary structure [393]. Dynamic programming was introduced in order to improve accuracy of the alignment [370, 394] as well as neighbour-joining methods to produce a more accurate guide tree [395]. An iterative refinement option was later introduced, which wherein individual sequences are removed from the alignment and realigned. A weights sum of pairs score is used to assess the alignment and recalculated after each iterative alignment [396].

6.1.2.2 MAFFT

MAFFT was developed with the goal of minimising computational time involved in large-scale MSAs, while still retaining the accuracy of the alignment [397]. The alignment algorithm is optimised to align residues based on volume and polarity, under the assumption that protein evolution favours substitutions which retain similar physicochemical properties of this nature, as described by Miyata *et al*. [398]. A fast Fourier transform (FFT) conversion is used to represent each amino acid in the sequence as a vector, describing its volume and polarity. Homologous regions between two proteins in an alignment are identified based on these values. The algorithms used by MAFFT perform up to 100 times faster than other widely used MSA programs. MAFFT has developed a series of different alignment strategies for specific types of data sets [399].

6.1.2.3 MUSCLE

MUSCLE uses a progressive alignment algorithm to produce MSAs with comparable accuracy to MAFFT [400, 401]. Two sets of progressive alignments are used. The first is based on a guide tree constructed using the sequences before they are aligned. After the initial alignment stage, a more accurate guide tree is constructed using this alignment and used for the next alignment stage. The alignment is then further refined using a profile function [400, 401].

6.1.2.4 PROMALS3D

PROMALS3D is a program that takes both sequence and structure into account when aligning two sequences [327]. The program uses both sequence- and structure-based constraints to produce an alignment. The sequence-based constraints use alignment information from PSI-BLAST and secondary structure predictions from PSIPRED to create a profile hidden Markov model (HMM) [356]. Homologous structures are identified using PSI-

BLAST or specified by the user and form part of the structure-based constraints. Using both sets of data has shown to greatly improve the accuracy of a sequence alignment [327].

6.1.2.5 T-Coffee, 3D-Coffee and Expresso

The Tree-based Consistency Objective Function for alignment Evaluation (T-Coffee) combines local and global alignment strategies with a progressive alignment to balance accuracy with speed [402]. It uses two pairwise-alignment libraries, one constructed using the Needleman & Wunsch [403] algorithm (global alignment library) and one using the SIM [404] algorithm (local alignment library) [371]. The alignments in each two libraries are scored based on sequence identity and combined into a single library, which balances the weighting from both local and global sequence alignment. This precompiled library is then used to guide the MSA, using a progressive alignment technique [402]. This program was improved by including structural information and released as 3DCoffee [371, 372]. To make the identifications of structures easier for users, Expresso was released, which runs a BLAST search against the PDB to automatically identify structural homologues to be used by 3DCoffee [405].

6.1.3 Model quality assessment programs

6.1.3.1 PROCHECK

PROCHECK is a suite of programs that consider only the stereochemistry when assessing a protein structure [406]. The suite runs five different programs and was designed to assess structures submitted to the PDB. As such it produces a number of different plots which allow a user to compare their structure to other structures with similar resolution and also provides an account for each amino acid in a structure, indicating which may be problematic. Though no longer maintained [407], the PROCKECK suite still forms part of the assessment of process of modern day homology modelling projects (e.g. [408, 409]).

6.1.3.2 VERIFY3D

Verify3D evaluates protein models using a set of three different environments considered to be favoured each amino acid [330]. These consider 1) How buried the side chain on a residue is; 2) The fraction of the residue exposed to polar contact and 3) The local secondary structure where the residue resides. By combining these factors, 18 different environments were assigned and located in proteins of known structure. Scores were assigned to each amino acid based on the degree to which the environmental class was either favoured or disfavoured by that amino acid. The Verify3D program sorts each residue in the model into one of these environmental classes and scores it accordingly. The sum of these scores is used to assess global model quality. The local quality of the model can be assessed by individual scores given to each residue. For graphical purposes, this score is often averaged over a window of 21 residues [330, 410].

6.1.3.3 ProSA

The Boltmann's principle is a knowledge-based description of the forces that stabilise a protein in solution [411]. Force fields derived from the Boltmann's principle were used to develop the Protein Structure Analysis (ProSA) model evaluation program [331, 332]. These consider protein-solvent interactions, as well as the energy inter-atomic interactions within the protein, expressed as a function of spatial separation [411]. The ProSA application constructs a polyprotein, approximately 50 000 residues in length, using fragments from known protein structures. This is constructed using only backbone atoms and includes atomic distances and steric information to ensure that all parts of the polyprotein are in a reasonable conformation. The model to be assessed is hidden within this structure, which is then assessed using a hide and seek method [331]. Here the structure is assessed in fragments of the same length of the model, which is used to establish a Z-score, indicating the extent to which the mean force potential of the model deviates from the random conformations of the

polyprotein [331]. The current online version of ProSA graphically represents the Z-score for the model in the context of all known structures in the PDB as a function of sequence length [332]. If the model Z-score falls outside the range established by native structures it is considered to be inaccurate. Additionally, the mean force potential of the model is plotted as a function of its amino acid sequence, averaged across both 10 and 40 residue windows [332].

6.1.3.4 MetaMQAPII

MetaMQAPII was developed to provide a means of accurately assessing the local structural quality of a model, as this is where other programs perform relatively poorly [329]. MetaMQAPII incorporates eight other model quality assessment programs, including Verify3D, ProSA, ANOLEA [412], BALA-SNAPP [413], TUNE [414], REFINER [415] and PROQRES [416]. Each residue in a structure is assessed by placing it into one of 315 groups, determined by assessing the structures submitted to the CASP 5 and CASP6 experiments, in the category of template-based modelling. First, the models were assessed using ProQ [416] and given a global accuracy score. Based on these scores, each model was divided equally into one of seven bins – Bin 1 containing residues from the worst-scored models, while bin 7 contained residues of models with the best ProQ scores. Secondly, in each bin, residues were further divided equally into another one of five bins, based on the extent to which they were buried within their respective structures, as determined by ResDepth. After this, each residue was further subdivided into one of three bins, based on residue type – hydrophobic, hydrophilic and other. Finally, each residue was then further subdivided into another one of three bins, based on local secondary structure. A combination of all these bins yielded 315 groups into which a residue of a structure could be divided. Linear regression model were developed to predict the RMSD of a residue in each group from its location in the native structure, using the scores of the combination of eight different MQAPs as input. Due to the

nature of linear regression models a scores from 0 – 100 is predicted, which is then converted to an RMSD value. MetaMQAPII returns a predicted GDT_TS score, which gives a prediction of the global quality of the model, as well as a predicted global RMSD value, describing the predicted deviation in angstroms from the true protein structure. Additionally, the residue scores of each MQAP used in the assessment is returned. Finally, a PDB coordinate file of the model is returned, wherein the B-factor for each residue is replaced with the ranking score from the meta-predictor calculation. This can be used to give a graphical representation of the model and easily identify problematic regions [329].

6.1.3.5 DOPE score

The Discrete Optimized Protein Energy (DOPE) score is a statistical potential derived using calculated PDFs of non-redundant structures from the PDB [417]. The score is dependent on inter-atomic distances of all atom pairs within a structure. Benchmark studies performed by Shen & Sali [417] indicated that the DOPE score function was better at finding the top model in a model set compared to other scoring functions. As the DOPE score is dependant of the length of the protein model, a normalised DOPE score (DOPE Z-score) was developed, which takes the length of the model into account [383].

6.1.4 Automated modelling tools

6.1.4.1 SWISS-Model

SWISS-Model is the longest-standing automated modelling server and focuses on being a user friendly interface for non-expert users to produce models for their research [418]. The server uses its own annotated template library, with structures from the PDB. Modelling is performed using ProMod-II and assessed using Qmean [419]. In 2005, SWISS-Model was criticised for lack of reliability, crashing more than 10% of the time it was run in bench-mark

tests [374]. This problem has been addressed by using MODELLER to perform modelling when ProMod-II crashes [418].

6.1.4.2 I-TASSER

The iterative threading assembly refinement (I-TASSER) server was the top performing automated modelling server in the CASP7 experiments [420]. In CASP assessments, I-TASSER is entered as "Zhang-Server" and even in the most recent round, CASP10, it performed best of the template-based protein structure predictors [421]. The principle used by I-TASSER is to combine threading and fragment assembly to produce models. The server integrates a number of different techniques and external applications.

6.1.4.3 YASARA

This is an application named "Yet Another Scientific Artificial Reality Application" (YARARA). The force-field, NOVA, was developed to addressed problems identified in CASP4, wherein structural refinement of protein models brought them further away from the native state [422]. YASARA uses molecular trees, which define the environment of each atom in a chemical structure, including topology information and bond types. The tree is constructed using all atoms, usually within three bonds of the reference atom. After identifying the chemical environment of a given atom in the structure, a force-field is selected from the reference set of molecular trees. The reference trees were constructed using the top 25 X-ray crystal structures from the PDB, based on resolution and ensuring less than 30% sequence identity between any two structures [422]. The YASARA algorithm was praised as one of the best approaches for structural refinement seen in CASP8 [423].

6.1.4.4 RosettaCM

RosettaCM describes a new approach to improve comparative modelling using Rosetta [196]. Using a threading approach, the target sequence is used to make a partial model for each of a

number of different templates. Full models are created using conformational sampling, using Rosetta, with structural restraints creating using information from the templates. Entered in CASP10 as "BAKER-ROSETTASERVER", this program performed best structural refinement [196] and produced the best model for the most challenging targets [421].

6.1.5 Proposed work

While automated modelling servers have advanced a great deal in recent years, there are still a number of inherent problems with using these. The most prominent of which is that the user does not know what is happening in the modelling process, nor do they have the ability to adjust parameters to achieve a better model. The aim of this chapter is to develop a set of scripts that automate the homology modelling process, when using MODELLER. The resulting automated modelling tool will need to be able to help users identify and evaluate templates, perform alignments, modelling and model evaluation. The use of the scripts will also be assessed using a simple benchmarking test. The eventual goal of this work is to provide an easy-to-use interface for users to reliably model protein targets. The tool aims to be transparent, allowing users to know what is happening in the background and enables them to exercise a degree of control over this process. The tool will also aim to allow modelling results to be reproducible.

6.2 Methodology

The automated homology modelling script has been written as a Python class, named Auto_model. This uses five other classes, which were written for this purpose and include 1) a PDB parser; 2) an HHPred 'runner'; 3) An alignment class; 4) a Pir file class; and 5) a MODELLER template class. The Auto_model script, as well as scripts used by Auto_model are given in Appendix H.

6.2.1 PDB parser class

This parsers files in PDB format and extracts information relevant to modelling, including SEQRES sequence; missing residues and the amino acid sequence, based on the coordinates section of the file. Also contains functionality to produce a PDB file containing the coordinates of a single chain from the original PDB file, as well as a method to check the alignment of a PIR file, by comparing the template sequence to the structure.

6.2.2 HHPred class

The HHPred class runs HHBlitz and HHSearch, which were installed as part of the HH-suite, version 2.0.16. This class also parses the results from HHSearch and stores information regarding the various templates found, including identifiers, query coverage, sequence identity matches, scores and E-values. This can also be used to extract alignment information between the target and template sequences and save them as in Fasta file format.

6.2.3 Sequence alignment class

This is used to run different alignment programs. Currently, these include MAFFT version 7.127b, MUSCLE version 3.8.31, Clustal Omega version 1.0.3 and T-Coffee version 10.00.r1613.

6.2.4 PIR file class

This is used to read and produce both Fasta files and PIR files. This class is used in combination with the PDB parser class to prepare PIR files for modelling.

6.2.5 MODELLER template class

This contains a shell with code common to modelling jobs run using MODELLER and is used to create and run modelling scripts for MODELLER. Additionally, this is used to specify different modelling parameters.

6.2.6 Automodel script

It consists of a number of functions that use the classes above to perform homology modelling using MODELLER. The only required input is a Fasta file containing the target sequence to be modelled. Optional input includes the names of template PDB files to be used for modelling and the name of an alignment file. The location of a PDB files directory, as well as an uploads and results directory can be specified. The purposes of these will be explained with respect to the algorithm (Figure 6.1).

6.2.6.1 Get templates step

If a template is not specified, this will use the HHPred class to find potential template structures. The program can be paused at this point in order to review the potential templates found, but by default the single top-scoring template returned by HHSearch is selected. Once the templates are selected, the script checks the PDB files directory for these. Those not found in the directory are downloaded from the PDBe, using a wget request – URL: http://www.ebi.ac.uk/pdbe-srv/view/files/[PDB ID]. The PDB parser is then used to make a PDB object for each template.

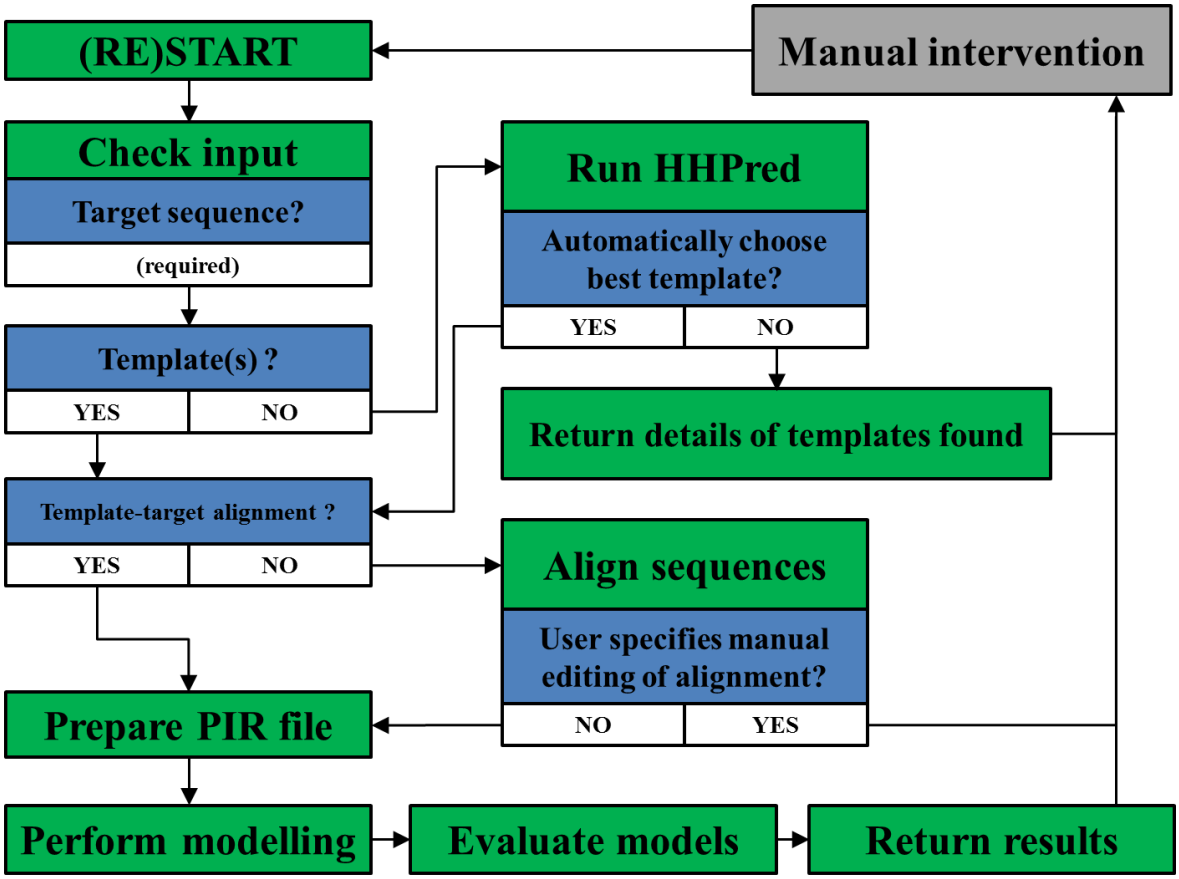


Figure 6.1: Algorithm used by automatic modelling script. The algorithm is explained in text. Stages are coloured as follows. Green: computational process; Blue: decision; white: outcome; Grey: manual process.

6.2.6.2 Alignment step

This step is skipped if an alignment is already provided. Otherwise, it simply calls the sequence alignment class to align the target sequence to the sequence(s) of the template(s). By default MAFFT-homologs is used to perform the alignment.

6.2.6.3 PIR file preparation step

This uses the alignment and the template PDB objects created using the PDB parser. Information for the PDB objects is used to provide start and end chains and residues in the PIR file and ensure that only residues present in the coordinate file are included in the alignment. The PIR file sequences are also trimmed at the N- and C-terminals so that these sections do not extend beyond the boundaries of the template sequence(s).

6.2.6.4 Modelling step

The MODELLER template class is used to create a modelling file so that MODELLER can be used to model the target protein of interest. The version of MODELLER available on the machine running the script can be specified as well the number of models to be produced and the level of refinement performed during modelling.

6.2.6.5 Evaluation step

This step calls MODELLER to assess the structures produced, as well as the templates used for modelling, using the normalized DOPE score (DOPE Z-score). The models are ranked according to this score and printed to a text file.

6.2.7 Preliminary benchmarking studies

6.2.7.1 Modelling known PDB structures using different alignment programs

The protocol for the benchmarking test is illustrated in Figure 6.2. David Brown from RUBi downloaded the entire PDB as part of his own project and kindly agreed to let me use these PDB files to perform benchmark studies. From this collection 100 different structures were selected at random and their sequences were sent to run through HHBlitz and HHSearch in order to find templates to attempt to model these targets. The top five suggested templates were recorded and binned according to their sequence identity, compared to their targets. From each bin, the suggested templates for 15 unique targets (i.e. one target structure was only binned in a single identity range). For each target-template grouping in each bin, the automated homology modelling protocol was run (Section 6.2.6, without manual intervention), using each of the four different sequence alignment options; MAFFT, MUSCLE, Clustal and Expresso. The DOPE Z-score was calculated for each template, target and model. The RMSD between the 1) template and target and 2) the top model and target

was calculated using BioPython. An RMSD difference value was calculated by subtracting the value of (1) from that of (2).

6.2.7.2 Remodelling of PDB files

Each target PDB file from 6.2.7.1 was also remodelled using its own structure as a template. The sequence used for modelling only considered the residues that had been experimentally solved (i.e. missing residues from the template file were left as gaps).

6.2.8 Web interface to the homology modelling tool

An online interface was designed by David Brown from RUBi, which uses the scripts described above to allow users to access and use the homology modelling tool via the internet.

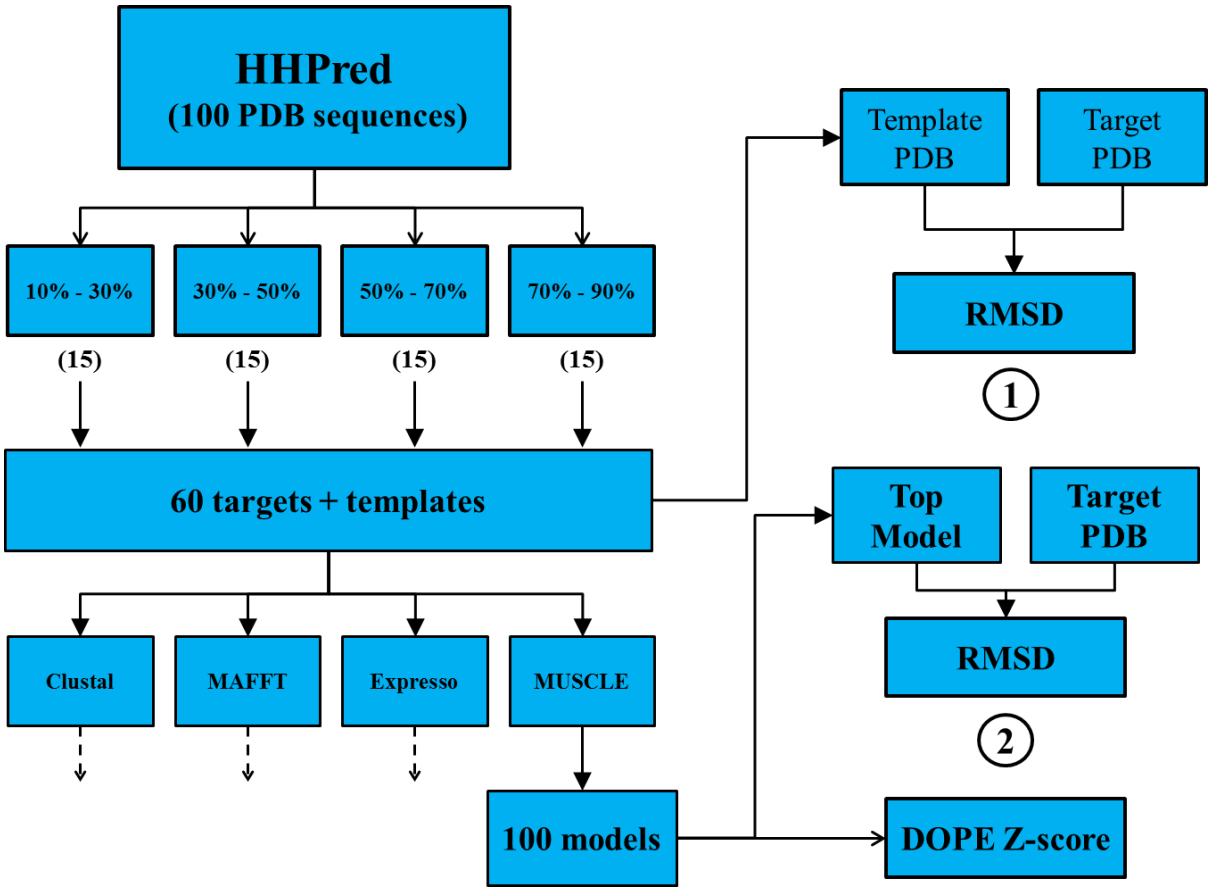


Figure 6.2: Benchmark test protocol. The algorithm followed to benchmark the automated modelling scripts is shown and discussed in more detail text.

6.3 Results and Discussion

6.3.1 Benchmarking tests

Apart from assessing the performance of the modelling scripts as a whole, the benchmarking tests were used to find and resolve problems found in each component of the tool. These are discussed in Sections 6.3.2 and 6.3.3. Due to the large-scale nature of the benchmark test, only the DOPE Z-score was used as an model quality score, since other MQAP servers such as MetaMQAPII [329] and ProSA [332] need to be accessed manually via their web interfaces. Benchmark studies performed by Shen & Sali [417] indicated that the DOPE score function was better at finding the top model in a model set compared to other scoring functions. The DOPE Z-scores of the different modelling sets are displayed in Figure 6.3. Both the template and target PDB files were included in this assessment in order to establish the DOPE Z-scores returned for crystal structures.

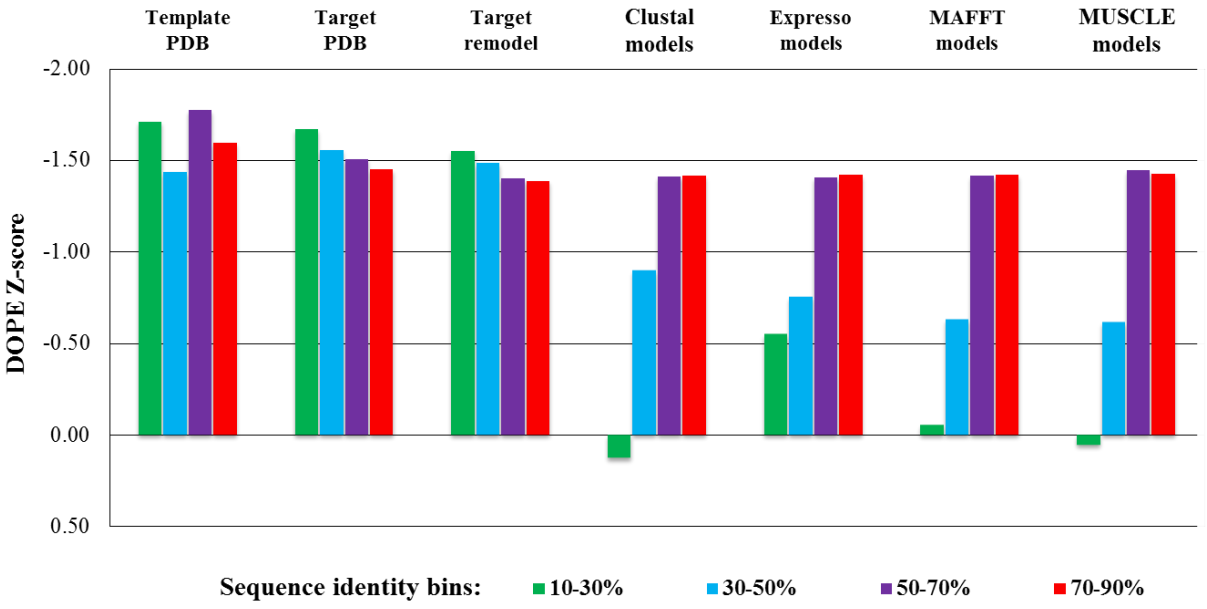


Figure 6.3: DOPE Z-scores calculated for templates and models from benchmarking test. Models are grouped based on the alignment program used to align the template to the target sequence. The template PDBs, target PDBs and remodelled targets are also shown. Groupings are further subdivided by the sequence identity between the templates and targets, coloured as indicated in the key. The DOPE Z-score axis is values are shown in reverse order, with negative values at the upper end of the axis, as these indicate better models.

In addition to this, the target PDB files were remodelled using their own crystal structures as templates, to represent modelling under ideal conditions (i.e. with an ideal template and a perfect sequence alignment). The MODELLER manual [383] states that a lower DOPE Z-score is favourable, with a score of less than -1.0 indicating a model is close to the native state of the protein. These values are reflected in the analysis performed on the PDB structures and remodelled targets. The models calculated for templates with greater than 50% sequence identity also fall within this range, regardless of sequence alignment program used. The differences in results produced by modelling using the various alignment programs were seen in the 10-30% and 30-50% sequence identity bins. Expresso alignments produced the best models when the sequence identity fell below 30%. This is probably because it is the only alignment program tested that uses structural information from the template when aligning sequences [405]. At sequence identity range of 30-50%, all alignment programs resulted in models with DOPE Z-scores below -0.5, which is promising.

Although DOPE Z-score is a good measure when predicting the quality models, it is still useful to measure the accuracy of modelling runs performed. Since target proteins with known structure were modelled, this can be achieved by measuring the RMSD between these targets and the models produced. To account for conformational differences between the template used for modelling and the target PDB files, an RMSD difference value was calculated. This indicates how much the model produced was closer to the target PDB, than the template used to model it was. The results of this, shown in Figure 6.4, were interesting. In the lower sequence identity ranges, where the template PDBs should be the most different to the targets, the models produced using Expresso were much closer to the target structure than the template. This was most evident in the 10-30% sequence identity range, which was also the range at which Expresso produced models with better measured DOPE Z-scores than the other alignment programs. These results, taken together, highlight the value of a good

sequence alignment when modelling. For the purposes of the automated homology modelling tool, this benchmark study indicates that at different sequence identity ranges, it may be better to align sequences using specific programs. Further studies will be required to firstly confirm this, but also to characterise the range so that a default program can be assigned.

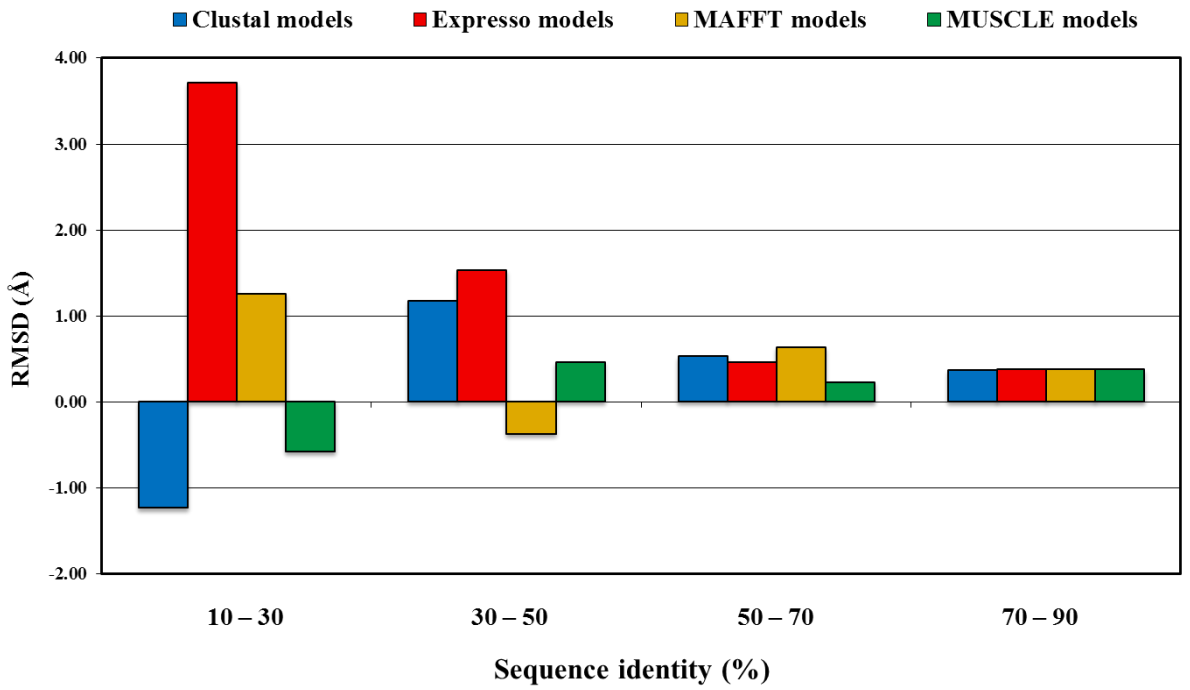


Figure 6.4: RMSD difference values calculated for templates and models from benchmarking test. Values are shown for models calculated and grouped according to the sequence identity between template and target, as well as the alignment program used, as shown in the figure key.

6.3.2 Alignment programs

The previous section showed the importance of the alignments when performing homology modelling; however, the alignment programs were the most common limitation when the benchmark was performed. Both MAFFT-homologues and Expresso require a connection to an external web server to find homologous sequences and structures. When running a relatively large number of alignments, this became problematic. The web servers that receive the request slowed down in response time with subsequent requests performed. After as few as five consecutive requests, the response time was long enough that the request was abandoned by the alignment program, resulting in the modelling run crashing. This problem

was fixed by enforcing a wait of five minutes between alignment runs that use these external web servers. If the automated modelling tool is to be incorporated into an online web service, then the exact wait required will need to be further characterised to ensure it is minimised and modelling runs proceed as smoothly as possible. Another problem was seen with Clustal. Sometimes, if the sequence identity was too low between template and target, the program crashed. As a result the models produced using other alignment programs could not be included in the analysis. The exact reason for this crash is not yet clear and will require further investigation.

6.3.3 Gaps in templates

Another problem identified through the benchmark studies was caused by gaps in template structures, which resulted from missing residues when the structure was solved. The resulting problem is specific to MODELLER, as the program attempts to bridge the gaps in the template structure. Sometimes it is impractical for MODELLER to do this, however, and it is better to simply model using the portions of the template that have been solved (i.e. leave the gap as is). The severity of this problem is dependant on the size of the gap. Using the missing residues information in the template PDB file, it is possible to determine this and alert the user to the potential problem.

6.3.4 Automated modelling web interface

The scripts used to create the homology modelling tool were used to create a web interface to perform homology modelling. This interface was designed by David Brown from RUBi and is presented in this thesis to show the application of the scripts described in this chapter. The web interface has been assigned the name “AutoModel”, and in this chapter will be referred to as such. The name is a contraction of the phrase “automated modelling”. However, due to this being the name of the MODELLER class which performs modelling, this name will

change in future, once the interface is ready for public use. The web front end to this tool (Figure 6.5) was built to reflect the functionality provided by the scripts.

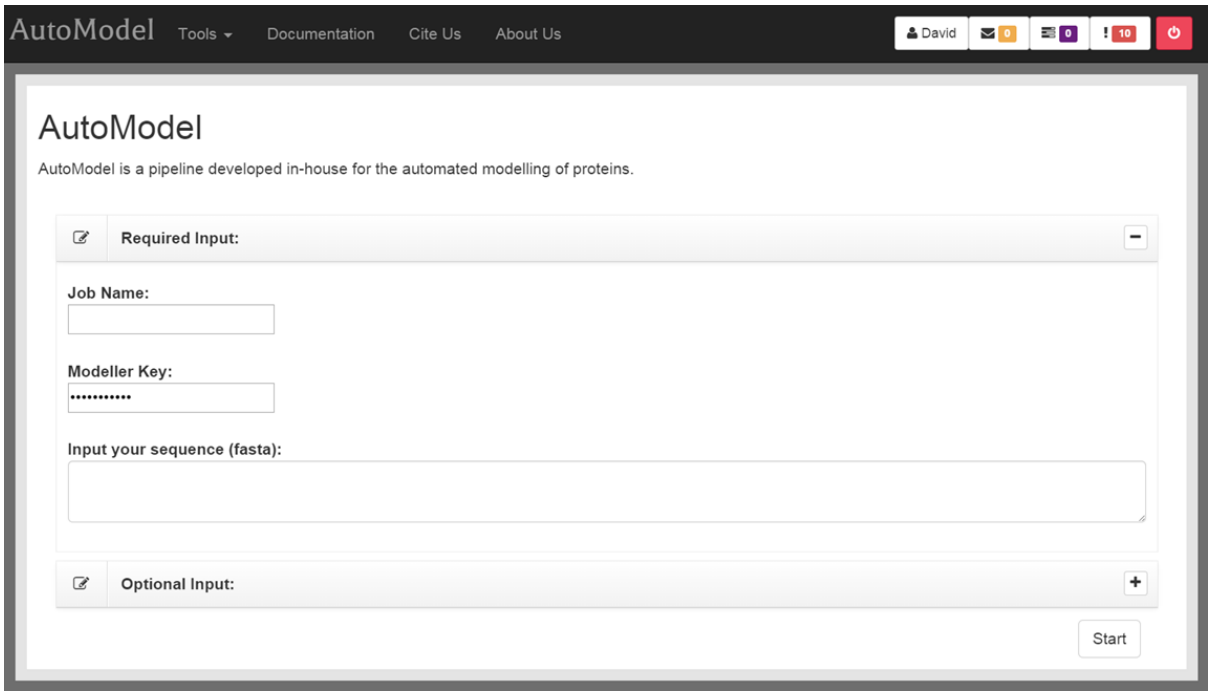


Figure 6.5: Current front end to the homology modelling tool. The interface to the AutoModel website is shown along with the required input. The optional input section is minimised.

At its core, the only required input is a sequence of the target protein in Fasta format. The site also requests a license key, which is required to ensure that MODELLER is used only used for academic purposes [383], and a job name, which allows users to keep track of their modelling runs and access the results at a later date. To allow users the freedom to take control of their homology modelling jobs, the optional input section is provided (Figure 6.6). This lists the various modelling options contained within the modelling scripts, as well as a job description section, which simply allows users to provide more information about their modelling job, for their own personal reference. The optional parameters are divided into three sections, including template identification, alignment options and parameters for MODELLER.

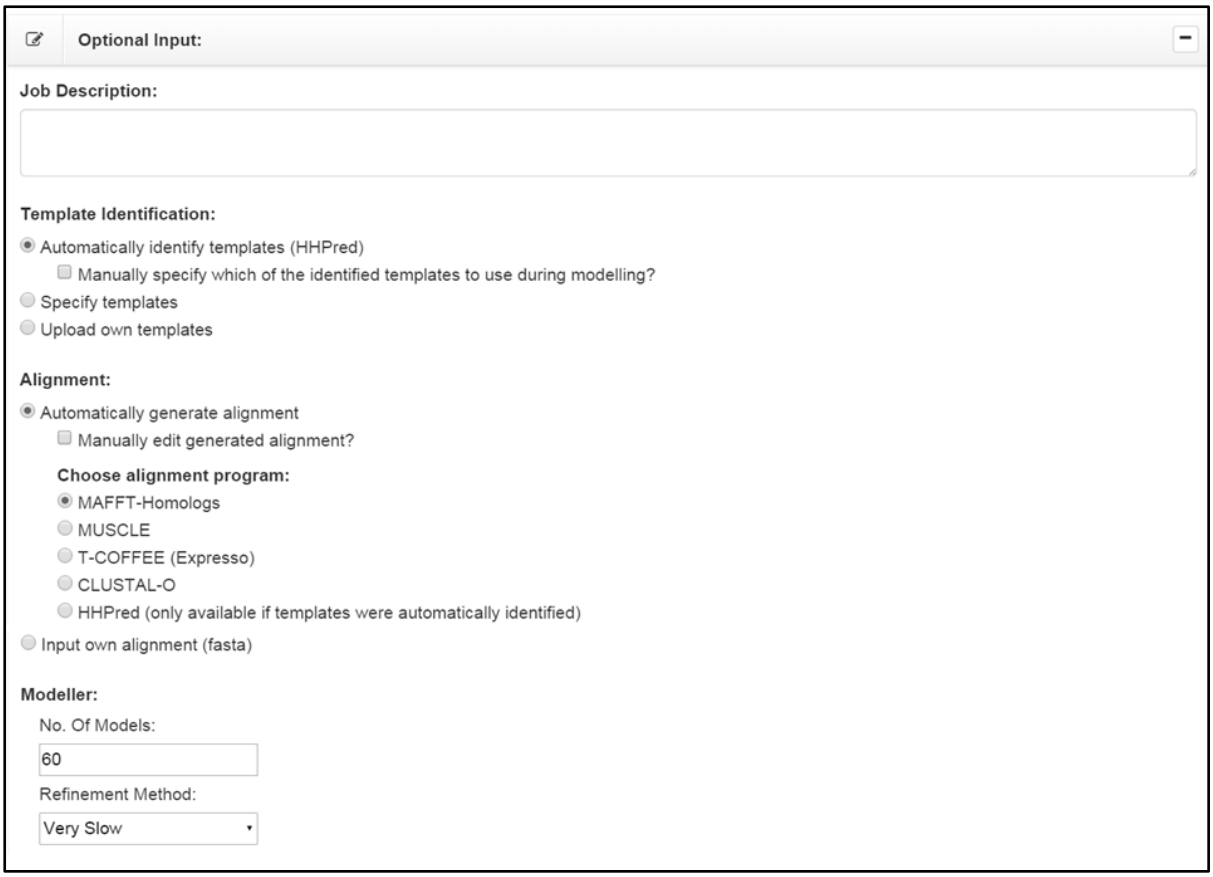


Figure 6.6: Optional input section of the AutoModel web interface. The optional input section, minimised in Figure 6.5, is displayed.

The layout of the optional input section largely follows the logic of the diagram represented in Figure 6.1 and is designed to accept user input where possible. For example, there is a check box provided for users to indicate whether or not they want to inspect and edit the alignment performed by the software provided. The page was also build using responsive design, meaning that the options provided will change, based on what is selected. An example of this is shown for the template identification section in Figure 6.7. If templates are selected to be identified by HHPred (Figure 6.7A), a check is box provided to allow users to review the templates returned by HHPred and choose the best one(s) rather than automatically using the template that was ranked best be HHPred. This tells the modelling scripts to halt and wait for user input. Alternatively, if users select the option allowing them to specify templates (Figure 6.7B), a text box appears for users to enter the PDB ID and chain

of the templates they wish to use for modelling. In this case, the scripts check if these PDB entries are in the PDB files directory, and, if not, send a wget request to retrieve these templates from the PDB, as in Section 6.2.6.1. Finally, users may wish to upload their own templates and select this option (Figure 6.7C), which simply adds the template file to the PDB files directory and continues with the modelling process as normal.

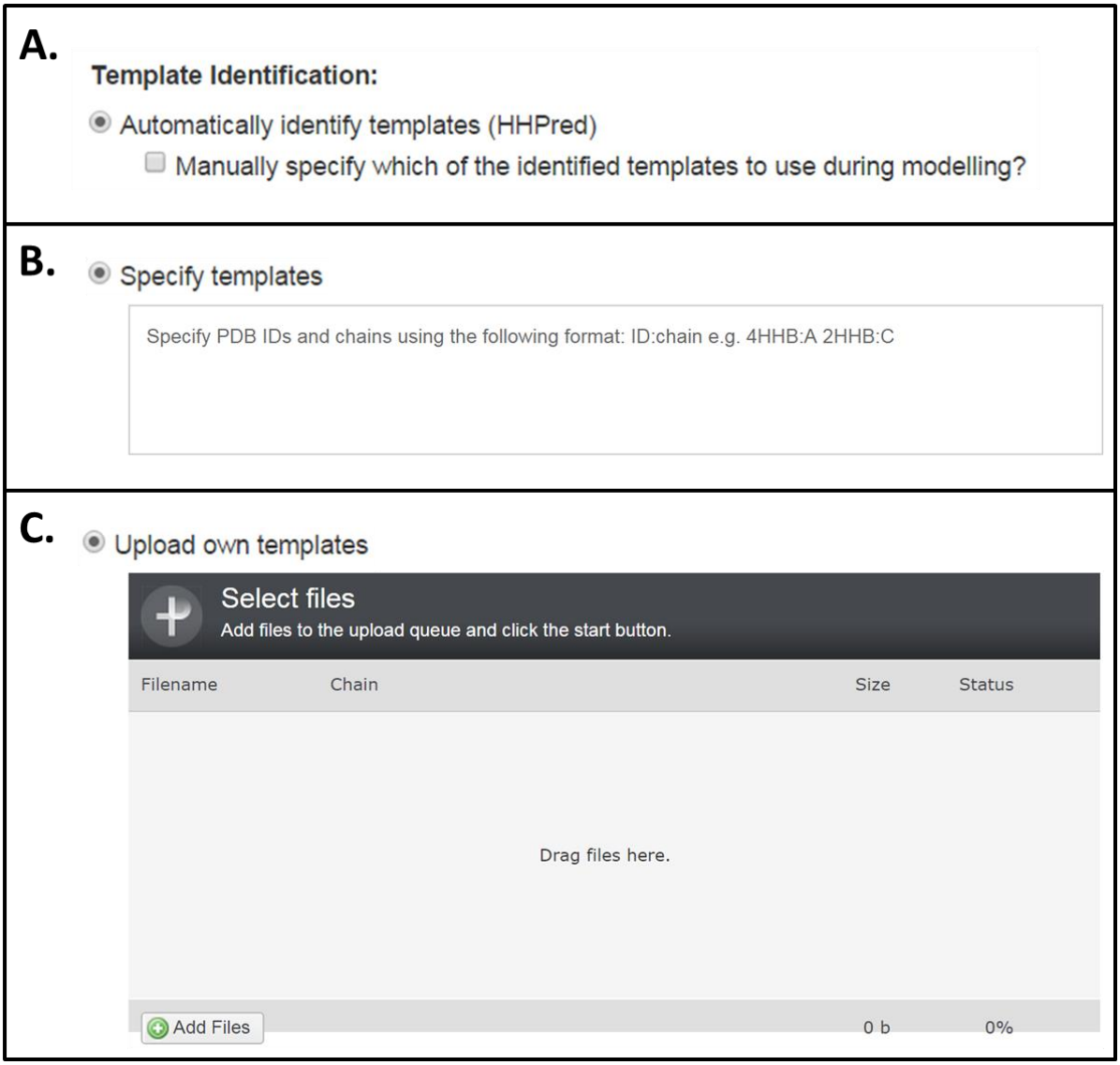


Figure 6.7: Responsive design of the option input section. The page is designed to display different content, depending on which options are selected and relevant to the input required. Panels A, B and C show the input types requested for each of the possible options in the templates section, which include a check box, a text box and a file upload widget, respectively.

Once the modelling parameters are set and modelling successfully completed, the web interface presents a summary of the job that was run (Figure 6.8A), as well as the results of each stage involved in the modelling process (Figure 6.8B-D).

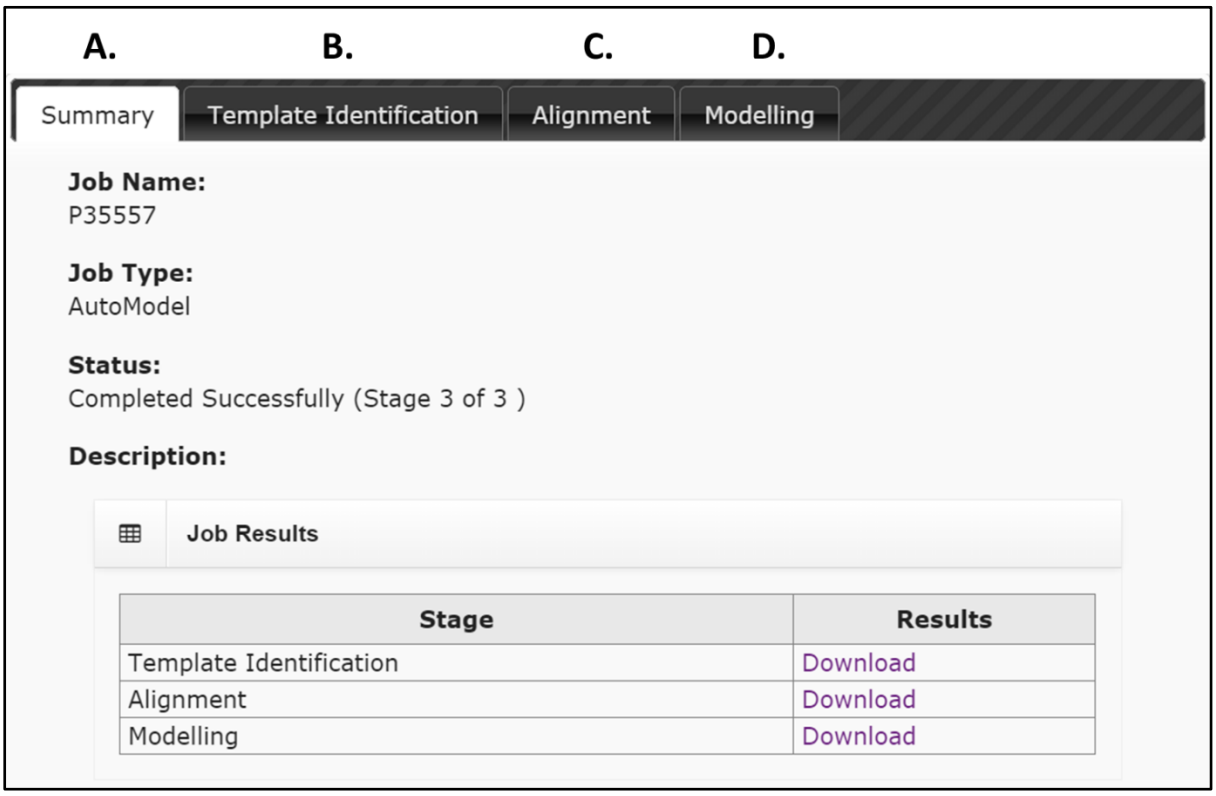


Figure 6.8: Results of a modelling job completed using the AutoModel web interface.

The modelling results section also includes a model evaluation segment, which can be used to visualise each model and assess their DOPE Z-scores (Figure 6.9). Users can download models with promising DOPE Z-scores and perform further model evaluation using external applications. The options selected in Figure 6.6 are the default parameters chosen for modelling - i.e. leaving this section as it is would be the same as just filling in the required input section and selecting “start”. The benefit of this is that it still allows the user to see what is happening behind the scenes in terms of what parameters were selected at each stage of modelling. For example, they know that HHPred was used to select the best template for

modelling and can refer to the specific input, which is saved in the job results 'summary' section (Figure 6.8).

Model	DopeZ
3hm8.ent_fit.pdb	-1.86205038456
P35557.B99990059.pdb	-1.6053584523
P35557.B99990059_fit.pdb	-1.60523488348
P35557.B99990021.pdb	-1.60315818967
P35557.B99990021_fit.pdb	-1.60305613781
P35557.B99990043_fit.pdb	-1.59990605519
P35557.B99990043.pdb	-1.59986670989
P35557.B99990005_fit.pdb	-1.58994739234
P35557.B99990005.pdb	-1.58989513687
P35557.B99990046_fit.pdb	-1.58797090607

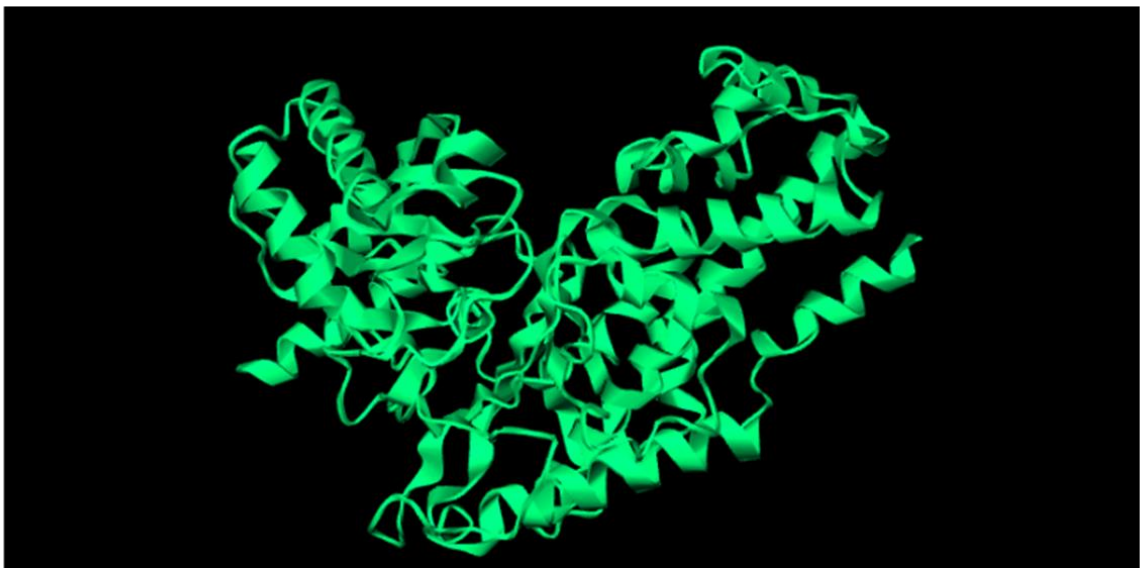


Figure 6.9: Model validation scores section of the AutoModel web interface. Within the results of for modelling stage (Figure 6.8D) is an option to view the validation scores, presented here. Models produced during the modelling process are tabulated and sorted by their DOPE Z-scores. Additionally, each model can be visualised using GLMol, as shown.

6.4 Conclusion

The automated modelling scripts were designed to provide a means to quickly assemble a modelling job and provide options to alter and improve modelling runs. The scripts can be

used to identify templates for modelling, perform alignments, modelling and preliminary model evaluation. The Z-dope scores also quickly identify the top models from these runs for further evaluation using external applications. MODELLER provides a range of functionality to perform homology modelling under different circumstances. One of the immediate future goals is to incorporate this functionality into the automated modelling scripts. Specifically, this includes complex modelling, modelling with ligands and modelling proteins with sulphide bridges. These will be provided to help users with less experience with these aspects of MODELLER. The scripts presented will also be incorporated into an online web application so that they may be used by anyone.

Chapter 7 Conclusions and future work

This thesis was divided into two sections and hence covered two broad aims. The first part aimed to establish an online database of South African NPs. This was divided into database construction, data collection and curation, followed by the creation of an online interface. The database was designed to be fully referenced, containing a broad range of information about each compound with potential to be expanded upon. A web interface was created to allow users to access this information. This was named SANCDB and can be accessed at https://sancdb.rubi.ru.ac.za/. Searches are provided to allow users to find compounds based on their names, structures, properties, source organisms, classifications, uses, as well as information pertaining to the referenced works used to retrieve this data. This information is also displayed in the compound summary page, provided for each compound entry in SANCDB. A RESTful API has been constructed and added to SANCDB, which allows information to be searched and downloaded from the database without having to access the web interface. This also enables information from SANCDB to be easily incorporated into external applications.

Various checks have been implemented to ensure the information is accurate and complete, including a multi-step process to allow depositors to make sure the compound SMILES uploaded matched the compound reported in literature. The compounds themselves have also been through error checking processes and 3D structures have been calculated using CORINA [113, 114], in combination with Open Babel [115] to convert these into different file formats for users to download and use in their own research. Compounds were further minimised using GAMESS [116], to provide 3D structures which are suitable for computational ligand docking experiments.

SANCDB currently contains information pertaining to 600 compounds housed within the database and linked to 170 different references. In terms of size, SANCDB is most similar to the Brazilian NP database, NuBBE [91], which houses 640 compounds. Additionally, there is the Central African library of natural compounds, ConMedNP [93], which contains 3200 compounds, but has no web interface, making these difficult to access. One of the immediate goals of SANCDB, moving forward, will be to increase the number of compounds contained within the database, while minimising overlap with other databases. ConMedNP covers compounds from around the Congo basin and does not extend to regions in Southern Africa. As such, the first step in extending the content of SANCDB will involve including compounds from Southern African countries. Secondly, SANCDB provides a web-based compound submission pipeline, developed by David Brown from RUBi. This will allow researchers with both published and unpublished compound data, not yet included in SANCDB, to deposit this information into the database. This will aim to provide a community-driven development of SANCDB content. Finally, the compound criteria for inclusion in SANCDB will be made broader to allow the database to grow. As mentioned with the submission pipeline, unpublished data can be included in the database, provided it recorded as such. These compounds can then be linked to the account of the user who supplied this information, rather than a reference. The second adjustment to the set of criteria used would be to include compound derivatives, rather than purely unmodified natural products. The inclusion of these will require additional information to be captured, specifically the process by which they were derived. The SANCDB interface will need to be modified to allow users to alter their search criteria include or exclude compounds derivatives, as well as the unpublished compounds. This would still link users to a greater quantity of NP chemical data, which suits their specific needs.

The second part of this thesis focused on protein structural studies. This initially involved the analysis of malarial proteins, specifically PfHsp70-x, PfHsp70-1, PfHsp90, PfHop and various PfJ proteins. The work and results presented formed part of two publications. One of these described the *is silico* characterisation of PfHsp70-x, including structural work to predict the interactions between PfHsp70-x and potential J protein cochaperones [207]. My contribution to the second paper [208] involved performing structural characterisation of the Hsp90:Hop convex interaction interface in both *P. falciparum* and human proteins. This was also compared to the concave interaction interface, which has been the target of many drug discovery efforts against Hop. Both sets of analysis involved similar techniques including the submission of modelled protein monomers to HADDOCK, protein-protein complex modelling, as well as analysis of these complexes using PIC and the Robetta Alanine Scanning web-server.

Results from the study involving the interaction of PfHsp70-x with potential J protein cochaperones yielded a number of interaction interfaces which agreed more closely with biochemical data from literature than the PDB structure of bovine Hsc70 cross-linked to auxillin used in complex modelling. Unfortunately, most of these poses appeared to be potentially valid. The Hsp90:Hop analyses indicated that the convex interaction sites were far less conserved between human and *P. falciparum* proteins compared to the concave sites. A number of hot spot residues were also identified specifically in the malarial complexes and not in their human counterparts. These residues may be potential site for drug development, aimed at specifically inhibiting the Hop:Hsp90 interaction of *P. falciparum*.

The final focus of this thesis was on the development of an automated homology modelling tool. The ideas incorporated into this tool were inspired by the work performed in the previous section, which involved a great deal of homology modelling. Also presented was the current interface for this tool, AutoModel, which demonstrates how the tool will be made

available to researchers. The preliminary benchmark tests performed for the homology modelling scripts were promising, but the tool does still require further development. The benchmarking tests help to identify bugs in the code or situations that have not yet been considered. The immediate plan for these is to upscale the number of targets to be modelled until the tool can reliably model proteins without any major errors. Once this has been achieved, the interface will be made ready for publication and made available at https://automodel.ru.ac.za/. The next step will be to expand on the functionality of the tool to perform the different types of modelling provided by MODELLER. The goals fro this work include making developments to the tool, such that it can be used for protein-protein complex modelling, protein-ligand modelled, as well as the modelling of proteins that have disulphide bridges.

The idea of tool development can be linked back to the SANCDB web interface. Apart from increasing the number of compounds in the database, the future of the website will lie in its tool development. Currently, the SMILES parser tool is under development. The most immediate goals for this software involve the generation of compound images from their SMILES, use in substructure searching and creation of compound fragments from a larger structure. The SMILES parser will be developed to enhance the functionality of the SANCDB website. The other aspect of tool development is the incorporation of the homology modelling tool into the SANCDB web interface. Since online compound databases are often used for computational screening of compounds, introducing the users of SANCDB to the homology modelling tool will become important, especially if they are unfamiliar with *in silico* modelling techniques. Figure 7.1 illustrates how the output from both SANCDB and the homology modelling tool form input for docking studies, indicating that although these separate sections of the thesis have their own merits and stand-alone applications, they can contribute to a common goal of *in silico* drug discovery.

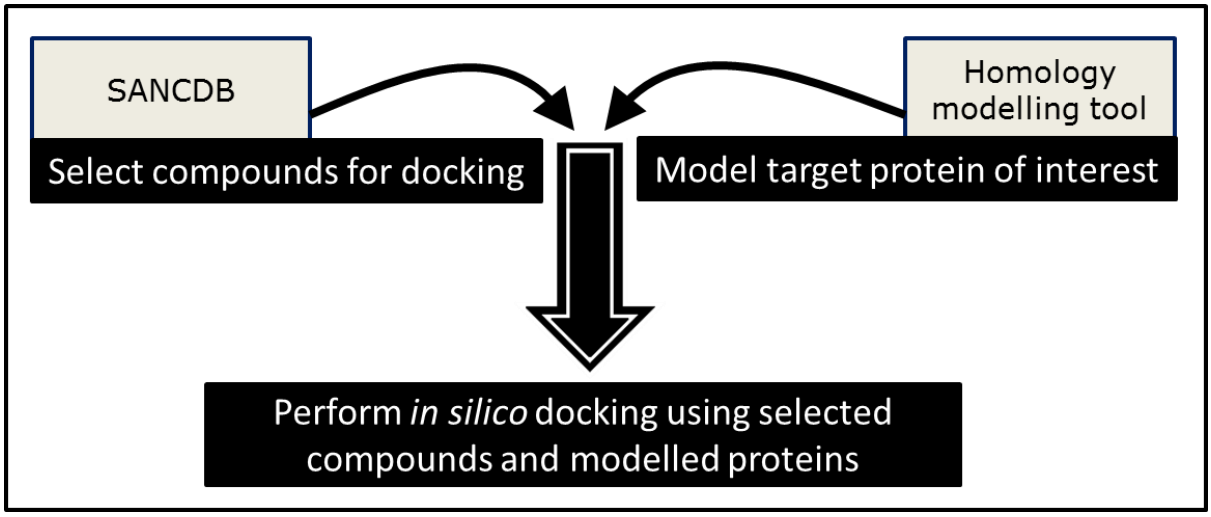


Figure 7.1: Simple schematic diagram, illustrating how the homology modelling tool can be used alongside SANCDB.

Chapter 8 References

1. Hanson JR: **The classes of natural product and their isolation**. In *Nat Prod Second Metab. Volume 17*. Edited by Abel EW.; 2003:1–34.

2. Pelay-Gimeno M, Tulla-Puche J, Albericio F: **"Head-to-side-chain" cyclodepsipeptides of marine origin.** *Mar Drugs* 2013, **11**:1693–1717.

3. Mishra BB, Tiwari VK: **Natural products: An evolving role in future drug discovery**. *Eur J Med Chem* 2011, **46**:4769–4807.

4. Paulsen IT, Brown MH, Skurray RA: **Proton-dependent multidrug efflux systems.** *Microbiol Rev* 1996, **60**:575–608.

5. Ncube NS, Afolayan AJ, Okoh AI: **Assessment techniques of antimicrobial properties of natural compounds of plant origin: current methods and future trends**. *African J Biotechnol* 2008, **7**:1797–1806.

6. Nascimento GGF, Locatelli J, Freitas PC, Silva GL: **Antibacterial activity of plant extracts and phytochemicals on antibiotic-resistant bacteria**. *Brazilian J Microbiol* 2000, **31**:247–256.

7. Cragg GM, Newman DJ: **Natural products: A continuing source of novel drug leads**. *Biochim Biophys Acta* 2013, **1830**:3670–3695.

8. Dias DA, Urban S, Roessner U: **A historical overview of natural products in drug discovery.** *Metabolites* 2012, **2**:303–336.

9. WHO: *WHO Traditional Medicine Strategy 2002-2005*. World Health Organization; 2013.

10. Stangeland T, Dhillion SS, Reksten H: **Recognition and development of traditional medicine in Tanzania**. *J Ethnopharmacol* 2008, **117**:290–299.

11. Freeman M, Motsei M: **Planning health care in South Africa - is there a role for traditional healers?** *Soc Sci Med* 1992, **34**:1183–1190.

12. Keji C, Hao X: **The integration of traditional Chinese medicine and Western medicine**. *Eur Rev* 2003, **11**:225–235.

13. Kasilo OM, Trapsida J-M: **Regulation of Traditional Medicine in the WHO African Region**. *African Heal Monit* 2010:25–31.

14. Kamsu-Foguem B, Diallo G, Foguem C: **Conceptual graph-based knowledge representation for supporting reasoning in African traditional medicine**. *Eng Appl Artif Intell* 2013, **26**:1348–1365.

15. WHO: *Progress Report on Decade of Traditional Medicine in the African Region*. 2011.

16. Light ME, Sparg SG, Stafford GI, Van Staden J: **Riding the wave: South Africa's contribution to ethnopharmacological research over the last 25 years**. *J Ethnopharmacol* 2005, **100**:127–130.

17. Louw CAM, Regnier TJC, Korsten L: **Medicinal bulbous plants of South Africa and their traditional relevance in the control of infectious diseases**. *J Ethnopharmacol* 2002, **82**:147–154.

18. Kelmanson JE, Jäger AK, Van Staden J: **Zulu medicinal plants with antibacterial activity**. *J Ethnopharmacol* 2000, **69**:241–246.

19. Tang J-L, Liu B-Y, Ma K-W: **Traditional Chinese medicine**. *Lancet* 2008, **372**:1938–1940.

20. Lukman S, He Y, Hui S-C: **Computational methods for Traditional Chinese Medicine: A survey.** *Comput Methods Programs Biomed* 2007, **88**:283–294.

21. Tariq A, Mussarat S, Adnan M: **Review on ethnomedicinal, phytochemical and pharmacological evidence of Himalayan anticancer plants**. *J Ethnopharmacol* 2015, **164**:96–119.

22. Steenkamp V, Gouws MC: **Cytotoxicity of six South African medicinal plant extracts used in the treatment of cancer**. *South African J Bot* 2006, **72**:630–633.

23. Jäger AK: **Is traditional medicine better off 25 years later?** *J Ethnopharmacol* 2005, **100**:3–4.

24. Schmitz R: **Friedrich Wilhelm Sertürner and the discovery of morphine.** *Pharm Hist* 1985, **27**:61–74.

25. Peters W: **Antimalarial drugs and their actions**. *Postgrad Med J* 1973, **49**:573–583.

26. Achan J, Talisuna AO, Erhart A, Yeka A, Tibenderana JK, Baliraine FN, Rosenthal PJ, D'Alessandro U: **Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria.** *Malar J* 2011, **10**:144.

27. Sneader W: **The discovery of aspirin: A reappraisal**. *Br Med J* 2000, **321**:1591–1594.

28. Jack DB: **One hundred years of aspirin**. *Lancet* 1997, **350**:437–439.

29. Ligon BL: **Penicillin: Its Discovery and Early Development**. *Semin Pediatr Infect Dis* 2004, **15**:52–57.

30. Wongsrichanalai C, Pickard AL, Wernsdorfer WH, Meshnick SR: **Epidemiology of drug-resistant malaria**. *Lancet Infect Dis* 2002, **2**:209–218.

31. Stermitz FR, Lorenz P, Tawara JN, Zenewicz LA, Lewis K: **Synergy in a medicinal plant: Antimicrobial action of berberine potentiated by 5'-methoxyhydnocarpin, a multidrug pump inhibitor**. *Proc Natl Acad Sci U S A* 2000, **97**:1433–1437.

32. Glaser KB, Mayer AMS: **A renaissance in marine pharmacology: From preclinical curiosity to clinical reality**. *Biochem Pharmacol* 2009, **78**:440–448.

33. Haefner B: **Drugs from the deep: Marine natural products as drug candidates**. *Drug Discov Today* 2003, **8**:536–544.

34. Martins A, Vieira H, Gaspar H, Santos S: **Marketed marine natural products in the pharmaceutical and cosmeceutical industries: Tips for success**. *Mar Drugs* 2014, **12**:1066–1101.

35. Lichtman MA: **A historical perspective on the development of the cytarabine (7days) and daunorubicin (3days) treatment regimen for acute myelogenous leukemia: 2013 the 40th anniversary of 7+3**. *Blood Cells, Mol Dis* 2013, **50**:119–130.

36. McIntosh M, Cruz LJ, Hunkapiller MW, Gray WR, Olivera BM: **Isolation and structure of a peptide toxin from the marine snail Conus magus**. *Arch Biochem Biophys* 1982, **218**:329–334.

37. Koski RR: **Omega-3-acid Ethyl Esters (Lovaza) For Severe Hypertriglyceridemia**. *Drug Forecast* 2008, **33**:271–303.

38. Gerwick WH, Moore BS: **Lessons from the past and charting the future of marine natural products drug discovery and chemical biology**. *Chem Biol* 2012, **19**:85–98.

39. Cuevas C, Pérez M, Martín MJ, Chicharro JL, Fernández-Rivas C, Flores M, Francesch A, Gallego P, Zarzuelo M, de La Calle F, García J, Polanco C, Rodríguez I, Manzanares I: **Synthesis of ecteinascidin ET-743 and phthalascidin Pt-650 from cyanosafracin B.** *Org Lett* 2000, **2**:2545–2548.

40. Hirata Y, Uemura D: **Halichondrins - antitumor polyether macrolides from a marine sponge**. *Pure Appl Chem* 1986, **58**:701–710.

41. Pettit GR, Kamano Y, Herald CL, Tuinman AA, Boettner FE, Kizu H, Schmidt JM, Baczynskyj L, Tomer KB, Bontems RJ: **The isolation and structure of a remarkable marine animal antineoplastic constituent: Dolastatin 10**. *J Am Chem Soc* 1987, **109**:6883–6885.

42. Duffy BC, Zhu L, Decornez H, Kitchen DB: **Early phase drug discovery: Cheminformatics and computational techniques in identifying lead series**. *Bioorganic Med Chem* 2012, **20**:5324–5342.

43. Baltz RH: **Marcel Faber Roundtable: Is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration?** *J Ind Microbiol Biotechnol* 2006, **33**:507–513.

44. Lam KS: **New aspects of natural products in drug discovery**. *Trends Microbiol* 2007, **15**:279–289.

45. Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DVS, Hertzberg RP, Janzen WP, Paslay JW, Schopfer U, Sittampalam GS: **Impact of high-throughput screening in biomedical research**. *Nat Rev Drug Discov* 2011, **10**:188–195.

46. Griffiths CL, Robinson TB, Lange L, Mead A: **Marine biodiversity in south africa: An evaluation of current states of knowledge**. *PLoS One* 2010, **5**:e12008.

47. Drewes SE: **Natural products research in South Africa : 1890-2010**. *S Afr J Sci* 2012, **108**:1–8.

48. Rabe T, Van Staden J: **Isolation of an antibacterial sesquiterpenoid from Warburgia salutaris**. *J Ethnopharmacol* 2000, **73**:171–174.

49. Salie F, Eagles PFK, Leng HMJ: **Preliminary antimicrobial screening of four South African Asteraceae species**. *J Ethnopharmacol* 1996, **52**:27–33.

50. Reddy L, Odhav B, Bhoola KD: **Natural products for cancer prevention: A global perspective**. *Pharmacol Ther* 2003, **99**:1–13.

51. Green E, Samie A, Obi CL, Bessong PO, Ndip RN: **Inhibitory properties of selected South African medicinal plants against Mycobacterium tuberculosis**. *J Ethnopharmacol* 2010, **130**:151–157.

52. Stafford GI, Pedersen ME, van Staden J, Jäger AK: **Review on plants with CNS-effects used in traditional South African medicine against mental diseases**. *J Ethnopharmacol* 2008, **119**:513–537.

53. Clarkson C, Maharaj VJ, Crouch NR, Grace OM, Pillay P, Matsabisa MG, Bhagwandin N, Smith PJ, Folb PI: **In vitro antiplasmodial activity of medicinal plants native to or naturalised in South Africa**. *J Ethnopharmacol* 2004, **92**:177–191.

54. Bessong PO, Obi CL, Andréola M-L, Rojas LB, Pouységu L, Igumbor E, Meyer JJM, Quideau S, Litvak S: **Evaluation of selected South African medicinal plants for inhibitory properties against human immunodeficiency virus type 1 reverse transcriptase and integrase**. *J Ethnopharmacol* 2005, **99**:83–91.

55. Van Vuuren SF: **Antimicrobial activity of South African medicinal plants**. *J Ethnopharmacol* 2008:462–472.

56. Kaido TL, Veale DJ, Havlik I, Rama DB: **Preliminary screening of plants used in South Africa as traditional herbal remedies during pregnancy and labour**. *J Ethnopharmacol* 1997, **55**:185–191.

57. Davies-coleman MT, Beukes DR: **Ten years of marine natural products research at Rhodes University**. *S Afr J Sci* 2004, **100**:539–544.

58. Blunt JW, Copp BR, Hu W-P, Munro MHG, Northcote PT, Prinsep MR: **Marine natural products.** *Nat Prod Rep* 2008, **25**:35–94.

59. Faulkner DJ: **Marine natural products**. *Nat Prod Rep* 2001, **18**:1–49.

60. Rishton GM: **Natural Products as a Robust Source of New Drugs and Drug Leads: Past Successes and Present Day Issues**. *Am J Cardiol* 2008, **101**:43D–49D.

61. Newman DJ, Cragg GM: **Natural products as sources of new drugs over the 30 years from 1981 to 2010**. *J Nat Prod* 2012, **75**:311–335.

62. Butler MS, Robertson A a B, Cooper M a: **Natural product and natural product derived drugs in clinical trials**. *Nat Prod Rep* 2014, **31**:1612–1661.

63. Ekins S, Mestres J, Testa B: **In silico pharmacology for drug discovery: Methods for virtual ligand screening and profiling**. *Br J Pharmacol* 2007, **152**:9–20.

64. Walker MJA, Barrett T, Guppy LJ: **Functional pharmacology: The drug discovery bottleneck?** *Drug Discov Today TARGETS* 2004, **3**:208–215.

65. Irwin JJ, Shoichet BK: **ZINC - A free database of commercially available compounds for virtual screening**. *J Chem Inf Model* 2005, **45**:177–182.

66. Ntie-Kang F, Mbah JA, Mbaze LM, Lifongo LL, Scharfe M, Hanna JN, Cho-Ngwa F, Onguéné PA, Owono Owono LC, Megnassan E, Sippl W, Efange SMN: **CamMedNP: Building the Cameroonian 3D structural natural products database for virtual screening**. *BMC Complement Altern Med* 2013, **13**:88.

67. Ntie-Kang F, Onguéné PA, Scharfe M, Owono Owono LC, Megnassan E, Mbaze LM, Sippl W, Efange SMN: **ConMedNP: A natural product library from Central African medicinal plants for drug discovery**. *RSC Adv* 2014, **4**:409.

68. Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, Mbah JA, Mbaze LM, Sippl W, Efange SMN: **AfroDb: A select highly potent and diverse natural product library from African medicinal plants**. *PLoS One* 2013, **8**:e78085.

69. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: A comprehensive resource for in silico drug discovery and exploration.** *Nucleic Acids Res* 2006, **34**(Database issue):D668–D672.

70. Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb KA, Tastan Bishop Özlem: **SANCDB: A South African Natural Compound Database**. *J Cheminform*, in press.

71. Buntrock RE: **Chemical registries - in the fourth decade of service.** *J Chem Inf Comput Sci* 2001, **41**:259–263.

72. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: A public information system for analyzing bioactivities of small molecules**. *Nucleic Acids Res* 2009, **37**:W623–33.

73. Wheeler DL, Barrett T, Benson D a, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R,

Tatusova T a, et al.: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2006, **34**(Database issue):D173–D180.

74. Li Q, Cheng T, Wang Y, Bryant SH: **PubChem as a public resource for drug discovery**. *Drug Discov Today* 2010, **15**:1052–1057.

75. Gaulton A, Bellis LJ, Bento a P, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res* 2012, **40**(Database issue):D1100–7.

76. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS: **DrugBank 3.0: A comprehensive resource for "Omics" research on drugs**. *Nucleic Acids Res* 2011, **39**(Database issue):D1035–D1041.

77. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, MacIejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS: **DrugBank 4.0: Shedding new light on drug metabolism**. *Nucleic Acids Res* 2014, **42**(Database Issue):D1091–D1097.

78. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG: **ZINC: A Free Tool to Discover Chemistry for Biology.** *J Chem Inf Model* 2012, **52**:1757–1768.

79. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK: **BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities**. *Nucleic Acids Res* 2007, **35**(Database issue):D198–D201.

80. Groom CR, Allen FH: **The Cambridge Structural Database in retrospect and prospect**. *Angew Chemie - Int Ed* 2014:662–671.

81. Degtyarenko K, De matos P, Ennis M, Hastings J, Zbinden M, Mcnaught A, Alcántara R, Darsow M, Guedj M, Ashburner M: **ChEBI: A database and ontology for chemical entities of biological interest**. *Nucleic Acids Res* 2008, **36**(Database issue):D344–D350.

82. Knox C, Shrivastava S, Stothard P, Eisner R, Wishart DS: **BioSpider: A web server for automating metabolome annotations.** *Pac Symp Biocomput* 2007, **12**:145–156.

83. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, MacInnis GD, et al.: **HMDB: The human metabolome database**. *Nucleic Acids Res* 2007, **35**(Database issue):D521–D526.

84. Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, Lee J, Liu P, Mandal R, Krishnamurthy R, Sinelnikov I, Wilson M, Wishart DS: **YMDB: The yeast metabolome database**. *Nucleic Acids Res* 2012, **40**(Database issue):D815–D820.

85. Frolkis A, Knox C, Lim E, Jewison T, Law V, Hau DD, Liu P, Gautam B, Ly S, Guo AC, Xia J, Liang Y, Shrivastava S, Wishart DS: **SMPDB: The small molecule pathway database**. *Nucleic Acids Res* 2009, **38**(Database issue):D480–D487.

86. Lim E, Pon A, Djoumbou Y, Knox C, Shrivastava S, Guo AC, Neveu V, Wishart DS: **T3DB: A comprehensively annotated database of common toxins and their targets**. *Nucleic Acids Res* 2009, **38**(Database issue):D781–D786.

87. Messer R, Fuhrer C a, Häner R: **Natural product-like libraries based on non-aromatic, polycyclic motifs**. *Curr Opin Chem Biol* 2005, **9**:259–265.

88. Ji ZL, Zhou H, Wang JF, Han LY, Zheng CJ, Chen YZ: **Traditional Chinese medicine information database**. *J Ethnopharmacol* 2006, **103**:501.

89. Shang S, Tan DS: **Advancing chemistry and biology through diversity-oriented synthesis of natural product-like libraries**. *Curr Opin Chem Biol* 2005, **9**:248–258.

90. Loub WD, Farnsworth NR, Soejarto DD, Quinn ML: **NAPRALERT: Computer handling of natural product research data**. *J Chem Inf Comput Sci* 1985, **25**:99–103.

91. Chen CYC: **TCM Database@Taiwan: The world's largest traditional Chinese medicine database for drug screening in silico**. *PLoS One* 2011, **6**:e15939.

92. Qiao X, Hou T, Zhang W, Guo S, Xu X: **A 3D structure database of components from Chinese traditional medicinal herbs**. *J Chem Inf Comput Sci* 2002, **42**:481–489.

93. Shen J, Xu X, Cheng F, Liu H, Luo X, Shen J, Chen K, Zhao W, Shen X, Jiang H: **Virtual screening on natural products for discovering active compounds and target information.** *Curr Med Chem* 2003, **10**:2327–2342.

94. Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X: **Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology**. *PLoS One* 2013, **8**:e62839.

95. Valli M, Dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD, Bolzani VS: **Development of a natural products database from the biodiversity of Brazil**. *J Nat Prod* 2013, **76**:439–444.

96. Lei J, Zhou J: **A Marine Natural Product Database**. *J Chem Inf Model* 2002, **42**:742–748.

97. Wiswesser WJ: **107 Years of Line-Formula Notations (1861-1968)**. *J Chem Doc* 1968, **8**:146–150.

98. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I: **InChI - The worldwide chemical structure identifier standard**. *J Cheminform* 2013, **5**:7.

99. Anderson E, Veith G, Weininger D: **SMILES: A Line Notation And Computerized Interpreter for Chemical Structures**. *US Environ Prot Agency, Environ Res Lab* 1987.

100. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J: **Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited**. *J Chem Inf Model* 1992, **32**:244–255.

101. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules**. *J Chem Inf Comput Sci* 1988, **28**:31–36.

102. Daylight Chemical Information Systems: **Daylight Theory Manual**. 2011:56.

103. Wagner A Ben: **SciFinder Scholar 2006: An empirical analysis of research topic query processing**. *J Chem Inf Model* 2006, **46**:767–774.

104. **CAS Fact Sheet** [http://www.cas.org/about-cas/cas-fact-sheets]

105. Bienfait B, Ertl P: **JSME: A free molecule editor in JavaScript**. *J Cheminform* 2013, **5**:24.

106. **SciFinder** [https://scifinder.cas.org]

107. Van Rossum G, de Boer J: **Interactive Testing Remote Servers Using the Python Programming Language**. *CWI Q* 1991, **4**:283–303.

108. **Django** [https://www.djangoproject.com/]

109. **MySQL** [https://www.mysql.com/]

110. **RU Theses,Rhodes University** [http://www.ru.ac.za/library/searchfind/theses/rutheses/]

111. **PubMed** [http://www.ncbi.nlm.nih.gov/pubmed]

112. **ScienceDirect** [www.sciencedirect.com]

113. **Google** [https://www.google.co.za/]

114. Filippov I V, Nicklaus MC: **Optical Structure Recognition Software To Recover Chemical Information: OSRA — An Open Source Solution**. *J Chem Inf Model* 2009, **49**:740–743.

115. **Depict** [http://www.daylight.com/daycgi/depict]

116. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open Babel: An Open chemical toolbox**. *J Cheminform* 2011, **3**:33.

117. **Indigo Tool Kit** [http://ggasoftware.com/opensource/indigo]

118. Sadowski J, Gasteiger J, Klebe G: **Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures**. *J Chem Inf Model* 1994, **34**:1000–1008.

119. Tetko I V, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin V a, Radchenko E V, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko V V: **Virtual computational chemistry laboratory - design and description**. *J Comput Aided Mol Des* 2005, **19**:453–463.

120. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su S, Windus TL, Dupuis M, Montgomery JA: **General atomic and molecular electronic structure system**. *J Comput Chem* 1993, **14**:1347–1363.

121. Lipinski C a, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings**. *Adv Drug Deliv Rev* 2001, **46**:3–26.

122. Elgorashi EE, Stafford GI, Van Staden J: **Acetylcholinesterase enzyme inhibitory effects of Amaryllidaceae alkaloids**. *Planta Med* 2004, **70**:260–262.

123. Warren FL: **The pyrrolizidine alkaloids. comparative structure, biological synthesis and pharmacology.** *Rec Chem Prog* 1958, **20**:1–21.

124. Warren FL: *The Pyrrolizidine Alkaloids. II*. Springer-Verlag; 1966.

125. Hastings J, Magka D, Batchelor C, Duan L, Stevens R, Ennis M, Steinbeck C: **Structure-based classification and ontology in chemistry**. *J Cheminform* 2012, **4**:8.

126. Amirkia V, Heinrich M: **Alkaloids as drug leads - A predictive structural and biodiversity-based analysis**. *Phytochem Lett* 2014, **10**:xlviii–liii.

127. Cushnie TPT, Cushnie B, Lamb AJ: **Alkaloids: An overview of their antibacterial, antibiotic-enhancing and antivirulence activities**. *Int J Antimicrob Agents* 2014, **44**:377–386.

128. Qiu S, Sun H, Zhang AH, Xu HY, Yan GL, Han Y, Wang XJ: **Natural alkaloids: Basic aspects, biological roles, and future perspectives**. *Chin J Nat Med* 2014, **12**:401–406.

129. Cao H, Chen X, Jassbi AR, Xiao J: **Microbial biotransformation of bioactive flavonoids**. *Biotechnol Adv* 2015, **33**:214–223.

130. Orhan DD, Özçelik B, Özgen S, Ergun F: **Antibacterial, antifungal, and antiviral activities of some flavonoids**. *Microbiol Res* 2010, **165**:496–504.

131. Mouradov A, Spangenberg G: **Flavonoids: A metabolic network mediating plants adaptation to their real estate**. *Front Plant Sci* 2014, **5**:1–16.

132. Cerella C, Dicato M, Diederich M: **Assembling the puzzle of anti-cancer mechanisms triggered by cardiac glycosides**. *Mitochondrion* 2013:225–234.

133. Ivanchina N V, Kicha AA, Stonik VA: **Steroid glycosides from marine organisms**. *Steroids* 2011, **76**:425–454.

134. Nising CF, Bräse S: **Highlights in steroid chemistry: Total synthesis versus semisynthesis**. *Angew Chemie - Int Ed* 2008, **47**:9389–9391.

135. Ajikumar PK, Tyo K, Carlsen S, Mucha O, Phon TH, Stephanopoulos G: **Terpenoids: Opportunities for biosynthesis of natural product drugs using engineered microorganisms**. *Mol Pharm* 2008, **5**:167–190.

136. Langenheim JH: **Higher plant terpenoids: A phytocentric overview of their ecological roles**. *J Chem Ecol* 1994, **20**:1223–1280.

137. Bobach C, Böhme T, Laube U, Püschel A, Weber L: **Automated compound classification using a chemical ontology**. *J Cheminform* 2012, **4**:1–12.

138. **Plantz Africa** [http://pza.sanbi.org/]

139. Kuroda M, Mimaki Y, Sashida Y: **Saundersiosides C-H, rearranged cholestane glycosides from the bulbs of Ornithogalum saundersiae and their cytostatic activity on HL-60 cells**. *Phytochemistry* 1999, **52**:435–443.

140. Sidwell WTL, Tamm C: **Errata: The homo-isoflavones II1). Isolation and structure of 4'-o-methyl-punctatin, autumnalin and 3,9-dihydro-autumnalin**. *Tetrahedron Lett* 1970, **11**:1578.

141. Koorbanally C, Crouch NR, Mulholland DA: **The phytochemistry and ethnobotany of the southern African genus Eucomis (Hyacinthaceae: Hyacinthoideae)**. In *Phytochem Adv Res. Volume 661*; 2006:69–85.

142. **Noel O'Blog: How to correct 3D coordinates at stereocenters** [http://baoilleach.blogspot.com/2009/10/how-to-correct-3d-coordinates-at.html]

143. Slee DH, Bhat AS, Nguyen TN, Kish M, Lundeen K, Newman MJ, McConnell SJ: **Pyrrolopyrazinedione-based inhibitors of human hormone-sensitive lipase**. *J Med Chem* 2003, **46**:1120–1122.

144. Timmer MSM, Verdoes M, Sliedregt L a JM, Van Der Marel GA, Van Boom JH, Overkleeft HS: **The Use of a Mannitol-Derived Fused Oxacycle as a Combinatorial Scaffold**. *J Org Chem* 2003, **68**:9406–9411.

145. Ivachtchenko A V, Tkachenko SE, Sandulenko YB, Vvedensky VY, Khvat A V: **A parallel solution-phase synthesis of substituted 3,7-diazabicyclo[3.3.1]nonanes.** *J Comb Chem* 2004, **6**:828–834.

146. Williams AJ: **A perspective of publicly accessible/open-access chemistry databases**. *Drug Discov Today* 2008, **13**:495–501.

147. Lipinski CA, Litterman NK, Southan C, Williams AJ, Clark AM, Ekins S: **Parallel Worlds of Public and Commercial Bioactive Chemistry Data**. *J Med Chem* 2015, **58**:2068–2076.

148. Williams AJ, Ekins S, Tkachenko V: **Towards a gold standard: Regarding quality in public domain chemistry databases and approaches to improving the situation**. *Drug Discov Today* 2012, **17**:685–701.

149. Radestock S: **Optimising chemical information workflows: Integrating Reaxys - use cases and applications**. *J Cheminform* 2013, **5**:P39.

150. Guthrie JP: **SAMPL4, a blind challenge for computational solvation free energies: The compounds considered**. *J Comput Aided Mol Des* 2014, **28**:151–168.

151. Granica S, Piwowarski JP, Czerwi ME, Kiss AK: **Phytochemistry, pharmacology and traditional uses of different Epilobium species (Onagraceae): A review**. *J Ethnopharmacol* 2014, **156**:316–346.

152. De Jesus NZT, de Souza Falcão H, Gomes IF, de Almeida Leite TJ, de Morais Lima GR, Barbosa-Filho JM, Tavares JF, da Silva MS, de Athayde-Filho PF, Batista LM: **Tannins, peptic ulcers and related mechanisms**. *Int J Mol Sci* 2012:3203–3228.

153. Pence HE, Williams A: **ChemSpider database**. *J Chem Educ* 2010, **87**:1123– 1124.

154. Williams A, Tkachenko V: **The Royal Society of Chemistry and the delivery of chemistry data repositories for the community**. *J Comput Aided Mol Des* 2014:1023–1030.

155. O'Boyle NM, Guha R, Willighagen EL, Adams SE, Alvarsson J, Bradley JC, Filippov I V, Hanson RM, Hanwell MD, Hutchison GR, James CA, Jeliazkova N, Lang ASID, Langner KM, Lonie DC, Lowe DM, Pansanel J, Pavlov D, Spjuth O, Steinbeck C, Tenderholt AL, Theisen KJ, Murray-Rust P: **Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on**. *J Cheminform* 2011, **3**:37.

156. Holovaty A, Kaplan-Moss J: *The Definitive Guide to Django: Web Development Done Right*. 2009.

157. Willard W: *HTML: A Beginner's Guide*. 4th edition. McGraw-Hill; 2009.

158. Schifreen BR: *The Web Book*. Oakworth Business; 2010.

159. Nixon R: *Learning PHP, MySQL and JavaScript*. 4th edition. O'Reilly Media; 2014.

160. Conery R, Haack P, Guthrie S, Hanselman S: *Professional ASP.NET MVC 1.0*. Wiley Publishing; 2009.

161. Laurie B, Laurie P: *Apache: The Definitive Guide*. 3rd edition. O'Reilly & Associates; 1999.

162. **Metis** [http://demo.onokumus.com/metis/]

163. **Molinspiration Cheminformatics** [http://www.molinspiration.com/]

164. Shehu A, Kavraki LE: **Modeling structures and motions of loops in protein molecules**. *Entropy* 2012, **14**:252–290.

165. Anfinsen CB: **Principles that govern the folding of protein chains.** *Science* 1973, **181**:223–230.

166. Terwilliger TC, Stuart D, Yokoyama S: **Lessons from Structural Genomics**. *Annu Rev Biophys* 2010, **38**:371–383.

167. Wlodawer A, Minor W, Dauter Z, Jaskolski M: **Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination**. *FEBS J* 2013, **280**:5705–5736.

168. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC: **A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis**. *Nature* 1958, **181**:662–666.

169. Smyth MS, Martin JHJ: **x Ray crystallography**. *J Clin Pathol Mol Pathol* 2000, **53**:8–14.

170. Leslie AGW: **The integration of macromolecular diffraction data**. *Acta Crystallogr Sect D Biol Crystallogr* 2006, **62**:48–57.

171. Frueh DP, Goodrich AC, Mishra SH, Nichols SR: **NMR methods for structural studies of large monomeric and multimeric proteins**. *Curr Opin Struct Biol* 2013, **23**:734–739.

172. Wüthrich K: **The way to NMR structures of proteins**. *Nat Struct Biol* 2001, **8**:923–925.

173. Moseley HN, Montelione GT: **Automated analysis of NMR assignments and structures for proteins.** *Curr Opin Struct Biol* 1999, **9**:635–642.

174. Inomata K, Ohno A, Tochio H, Isogai S, Tenno T, Nakase I, Takeuchi T, Futaki S, Ito Y, Hiroaki H, Shirakawa M: **High-resolution multi-dimensional NMR spectroscopy of proteins in human cells**. *Nature* 2009, **458**:106–109.

175. Takayama K, Nakase I, Michiue H, Takeuchi T, Tomizawa K, Matsui H, Futaki S: **Enhanced intracellular delivery using arginine-rich peptides by the addition of penetration accelerating sequences (Pas)**. *J Control Release* 2009, **138**:128–133.

176. Milne JLS, Borgnia MJ, Bartesaghi A, Tran EEH, Earl LA, Schauder DM, Lengyel J, Pierson J, Patwardhan A, Subramaniam S: **Cryo-electron microscopy - A primer for the non-microscopist**. *FEBS J* 2013, **280**:28–45.

177. Orlova E V, Saibil HR: **Structural analysis of macromolecular assemblies by electron microscopy**. *Chem Rev* 2011, **111**:7710–7748.

178. Jonic S, Vénien-Bryan C: **Protein structure determination by electron cryo-microscopy**. *Curr Opin Pharmacol* 2009, **9**:636–642.

179. Lučić V, Rigort A, Baumeister W: **Cryo-electron tomography: The challenge of doing structural biology in situ**. *J Cell Biol* 2013, **202**:407–419.

180. Zhang Y: **Protein structure prediction: When is it useful?** *Curr Opin Struct Biol* 2009, **19**:145–155.

181. Zhang Y: **Progress and challenges in protein structure prediction**. *Curr Opin Struct Biol* 2008, **18**:342–348.

182. Illergård K, Ardell DH, Elofsson A: **Structure is three to ten times more conserved than sequence - a study of structural response in protein cores.** *Proteins* 2009, **77**:499–508.

183. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng* 1999, **12**:85–94.

184. Di Luccio E, Koehl P: **A quality metric for homology modeling: the H-factor.** *BMC Bioinformatics* 2011, **12**:1–19.

185. Krieger E, Nabuurs SB, Vriend G: **Homology modeling**. In *Struct Bioinforma*. Edited by Bourne PE, Weissig H. Wiley-Liss; 2003:507–521.

186. Read RJ, Chavali G: **Assessment of CASP7 predictions in the high accuracy template-based modeling category**. *Proteins Struct Funct Bioinforma* 2007, **69**:27–37.

187. Joseph AP, Srinivasan N, De Brevern AG: **Progressive structure-based alignment of homologous proteins: Adopting sequence comparison strategies**. *Biochimie* 2012, **94**:2025–2034.

188. Torda AE: **Perspectives in protein-fold recognition**. *Curr Opin Struct Biol* 1997, **7**:200–205.

189. Buchete N V, Straub JE, Thirumalai D: **Development of novel statistical potentials for protein fold recognition**. *Curr Opin Struct Biol* 2004, **14**:225–232.

190. Fiser A: **Template-based protein structure modeling**. *Methods Mol Biol* 2010, **673**:73–94.

191. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J: **On the origin and highly likely completeness of single-domain protein structures**. *Proc Natl Acad Sci U S A* 2006, **103**:2605–2610.

192. Jauch R, Yeo HC, Kolatkar PR, Clarke ND: **Assessment of CASP7 structure predictions for template free targets**. *Proteins Struct Funct Bioinforma* 2007, **69**:57–67.

193. Klepeis JL, Floudas CA: **ASTRO-FOLD: A combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence**. *Biophys J* 2003, **85**:2119–2146.

194. Bradley P, Misura KMS, Baker D: **Toward high-resolution de novo structure prediction for small proteins**. *Science* 2005, **309**:1868–1871.

195. Simons KT, 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* 1997, **268**:209–225.

196. Song Y, Dimaio F, Wang RYR, Kim D, Miles C, Brunette T, Thompson J, Baker D: **High-resolution comparative modeling with RosettaCM**. *Structure* 2013, **21**:1735–1742.

197. Moult J: **Rigorous performance evaluation in protein structure modelling and implications for computational biology**. *Philos Trans R Soc Lond B Biol Sci* 2006, **361**:453–458.

198. Chruszcz M, Domagalski M, Osinski T, Wlodawer A, Minor W: **Unmet challenges of structural genomics**. *Curr Opin Struct Biol* 2010, **20**:587–597.

199. Grabowski M, Chruszcz M, Zimmerman MD, Kirillova O, Minor W: **Benefits of structural genomics for drug discovery research**. *Infect Disord Drug Targets* 2009, **9**:459–474.

200. Joachimiak A: **High-throughput crystallography for structural genomics**. *Curr Opin Struct Biol* 2009, **19**:573–584.

201. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C: **PSI-2: Structural Genomics to Cover Protein Domain Family Space**. *Structure* 2009, **17**:869–881.

202. Sharma P, Sharma S, Chawla M, Mitra A: **Modeling the noncovalent interactions at the metabolite binding site in purine riboswitches**. *J Mol Model* 2009, **15**:633–649.

203. Sliwoski G, Kothiwale S, Meiler J, Lowe E: **Computational Methods in Drug Discovery**. *Pharmacol Rev* 2014, **66**:334–395.

204. Budzik B, Garzya V, Shi D, Walker G, Woolley-Roberts M, Pardoe J, Lucas A, Tehan B, Rivero RA, Langmead CJ, Watson J, Wu Z, Forbes IT, Jin J: **Novel N-substituted benzimidazolones as potent, selective, CNS-penetrant, and orally active M1 mAChR agonists**. *ACS Med Chem Lett* 2010, **1**:244–248.

205. Morya VK, Dung NH, Singh BK, Lee H-B, Kim E: **Homology modelling and virtual screening of P-protein in a quest for novel antimelanogenic agent and In vitro assessments**. *Exp Dermatol* 2014, **23**:838–842.

206. Akhoon BA, Singh KP, Varshney M, Gupta SK, Shukla Y, Gupta SK: **Understanding the Mechanism of Atovaquone Drug Resistance in Plasmodium falciparum Cytochrome b Mutation Y268S Using Computational Methods**. *PLoS One* 2014, **9**:e110041.

207. Hatherley R, Blatch GL, Tastan Bishop Ö: **Plasmodium falciparum Hsp70-x: A heat shock protein at the host-parasite interface**. *J Biomol Struct Dyn* 2014, **32**:1766–1779.

208. Hatherley R, Clitheroe C, Faya N, Tastan Bishop Özlem: **Plasmodium falciparum Hop: detailed analysis on complex formation with Hsp70 and Hsp90**. *Biochem Biophys Res Commun* 2015, **456**:440–445.

209. WHO: *World Malaria Report 2014*. World Health Organization; 2014.

210. Fujioka H, Aikawa M: **Structure and life cycle**. In *Malar Immunol. Volume 80*. 2nd edition. Edited by Perlmann P, Troye-Blomberg M. Karger Publishers; 2002:1–26.

211. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ: **A proteomic view of the Plasmodium falciparum life cycle**. *Nature* 2002, **419**:520–526.

212. Li H, Child MA, Bogyo M: **Proteases as regulators of pathogenesis: Examples from the Apicomplexa**. *Biochim Biophys Acta - Proteins Proteomics* 2012, **1824**:177–185.

213. Lengeler C: **Insecticide-treated bed nets and curtains for preventing malaria**. *Cochrane database Syst Rev* 2009:CD000363.

214. Wilson AL, Dhiman RC, Kitron U, Scott TW, van den Berg H, Lindsay SW: **Benefit of Insecticide-Treated Nets, Curtains and Screening on Vector Borne Diseases, Excluding Malaria: A Systematic Review and Meta-analysis**. *PLoS Negl Trop Dis* 2014, **8**:e3228.

215. McGready R: **Intermittent preventive treatment of malaria in infancy**. *Lancet* 2009, **374**:1478–1480.

216. Aponte JJ, Schellenberg D, Egan A, Breckenridge A, Carneiro I, Critchley J, Danquah I, Dodoo A, Kobbe R, Lell B, May J, Premji Z, Sanz S, Sevene E, Soulaymani-Becheikh R, Winstanley P, Adjei S, Anemana S, Chandramohan D, Issifou S, Mockenhaupt F, Owusu-Agyei S, Greenwood B, Grobusch MP, Kremsner PG, Macete E, Mshinda H, Newman RD, Slutsker L, Tanner M, et al.: **Efficacy and safety of intermittent preventive treatment with sulfadoxine-pyrimethamine for malaria in African infants: A pooled analysis of six randomised, placebo-controlled trials**. *Lancet* 2009, **374**:1533–1542.

217. Gosling RD, Gesase S, Mosha JF, Carneiro I, Hashim R, Lemnge M, Mosha FW, Greenwood B, Chandramohan D: **Protective efficacy and safety of three antimalarial regimens for intermittent preventive treatment for malaria in infants: A randomised, double-blind, placebo-controlled trial**. *Lancet* 2009, **374**:1521–1532.

218. Thwing J, Eisele TP, Steketee RW: **Protective efficacy of malaria case management and intermittent preventive treatment for preventing malaria mortality in children: A systematic review for the Lives Saved Tool**. *BMC Public Health* 2011, **11**:S14.

219. Sinclair D, Zani B, Donegan S, Olliaro P, Garner P: **Artemisinin-based combination therapy for treating uncomplicated malaria**. *Cochrane Database Syst Rev* 2009:CD007483.

220. WHO: *Good Procurement Practices for Artemisinin-Based Antimalarial Medicines*. World Health Organization; 2010.

221. WHO: *Guidelines for the Treatment of Malaria*. 2nd edition. World Health Organization; 2010.

222. Akide-Ndunge OB, Tambini E, Giribaldi G, McMillan PJ, Müller S, Arese P, Turrini F: **Co-ordinated stage-dependent enhancement of Plasmodium falciparum antioxidant enzymes and heat shock protein expression in parasites growing in oxidatively stressed or G6PD-deficient red blood cells**. *Malar J* 2009, **8**:113.

223. Banumathy G, Singh V, Pavithra SR, Tatu U: **Heat shock protein 90 function is essential for Plasmodium falciparum growth in human erythrocytes**. *J Biol Chem* 2003, **278**:18336–18345.

224. Chiang AN, Valderramos JC, Balachandran R, Chovatiya RJ, Mead BP, Schneider C, Bell SL, Klein MG, Huryn DM, Chen XS, Day BW, Fidock DA, Wipf P, Brodsky JL: **Select pyrimidinones inhibit the propagation of the malarial parasite, Plasmodium falciparum**. *Bioorganic Med Chem* 2009, **17**:1527–1533.

225. Kumar R, Musiyenko A, Barik S: **The heat shock protein 90 of Plasmodium falciparum and antimalarial activity of its inhibitor, geldanamycin**. *Malar J* 2003, **2**:30.

226. Ramya TNC, Surolia N, Surolia A: **15-Deoxyspergualin modulates Plasmodium falciparum heat shock protein function**. *Biochem Biophys Res Commun* 2006, **348**:585–592.

227. Shonhai A: **Plasmodial heat shock proteins: Targets for chemotherapy**. *FEMS Immunol Med Microbiol* 2010, **58**:61–74.

228. Acharya P, Kumar R, Tatu U: **Chaperoning a cellular upheaval in malaria: Heat shock proteins in Plasmodium falciparum**. *Mol Biochem Parasitol* 2007, **153**:85–94.

229. Shonhai A, Botha M, de Beer TAP, Boshoff A, Blatch GL: **Structure-function study of a Plasmodium falciparum Hsp70 using three dimensional modelling and in vitro analyses**. *Protein Pept Lett* 2008, **15**:1117–1125.

230. Shonhai A, Boshoff A, Blatch GL: **The structural and functional diversity of Hsp70 proteins from Plasmodium falciparum**. *Protein Sci* 2007, **16**:1803–1818.

231. Robert J: **Evolution of heat shock protein and immunity**. *Dev Comp Immunol* 2003, **27**:449–464.

232. Tardieux I, Baines I, Mossakowska M, Ward GE: **Actin-binding proteins of invasive malaria parasites and the regulation of actin polymerization by a complex of 32/34-kDa proteins associated with heat shock protein 70kDa**. *Mol Biochem Parasitol* 1998, **93**:295–308.

233. Cowman AF, Crabb BS: **The Plasmodium falciparum genome - a blueprint for erythrocyte invasion**. *Science* 2002, **298**:126–128.

234. Subjeck JR, Sciandra JJ, Chao CF, Johnson RJ: **Heat shock proteins and biological response to hyperthermia**. *Br J Cancer Suppl* 1982, **5**:127–131.

235. Bukau B, Horwich AL: **The Hsp70 and Hsp60 chaperone machines**. *Cell* 1998:351–366.

236. Hartl FU, Hayer-Hartl M: **Molecular chaperones in the cytosol: From nascent chain to folded protein**. *Science* 2002, **295**:1852–1858.

237. Borges JC, Ramos CHI: **Characterization of nucleotide-induced changes on the quaternary structure of human 70 kDa heat shock protein Hsp70.1 by analytical ultracentrifugation**. *BMB Rep* 2009, **42**:166–171.

238. Vogel M, Mayer MP, Bukau B: **Allosteric regulation of Hsp70 chaperones involves a conserved interdomain linker**. *J Biol Chem* 2006, **281**:38705–38711.

239. Han W, Christen P: **Mutations in the interdomain linker region of DnaK abolish the chaperone action of the DnaK/DnaJ/GrpE system**. *FEBS Lett* 2001, **497**:55–58.

240. Swain JF, Dinler G, Sivendran R, Montgomery DL, Stotz M, Gierasch LM: **Hsp70 Chaperone Ligands Control Domain Association via an Allosteric Mechanism Mediated by the Interdomain Linker**. *Mol Cell* 2007, **26**:27–39.

241. Strub A, Zufall N, Voos W: **The Putative Helical Lid of the Hsp70 Peptide-binding Domain is Required for Efficient Preprotein Translocation into Mitochondria**. *J Mol Biol* 2003, **334**:1087–1099.

242. Wang T-F, Chang J-H, Wang C: **Identification of the Peptide Binding Domain of hsc7O**. *J Biol Chem* 1993, **268**:26049–26051.

243. Moro F, Fernández-Sáiz V, Muga A: **The Lid Subdomain of DnaK Is Required for the Stabilization of the Substrate-binding Site**. *J Biol Chem* 2004, **279**:19600–19606.

244. Zhu X, Zhao X, Burkholder WF, Gragerov A, Ogata CM, Gottesman ME, Hendrickson WA: **Structural Analysis of Substrate Binding by the Molecular Chaperone DnaK**. *Science* 1996, **272**:1606–1614.

245. Liberek K, Marszalek J, Ang D, Georgopoulos C, Zylicz M: **Escherichia coli DnaJ and GrpE heat shock proteins jointly stimulate ATPase activity of DnaK**. *Proc Natl Acad Sci U S A* 1991, **88**:2874–2878.

246. Karzai AW, McMacken R: **A bipartite signaling mechanism involved in DnaJ-mediated activation of the Escherichia coli DnaK protein**. *J Biol Chem* 1996, **271**:11236–11246.

247. Rüdiger S, Schneider-Mergener J, Bukau B: **Its substrate specificity characterizes the DnaJ co-chaperone as a scanning factor for the DnaK chaperone**. *EMBO J* 2001, **20**:1042–1050.

248. Höhfeld J, Minami Y, Hartl FU: **Hip, a novel cochaperone involved in the eukaryotic Hsc70/Hsp40 reaction cycle**. *Cell* 1995, **83**:589–598.

249. Harrison CJ, Hayer-Hartl M, Di Liberto M, Hartl F, Kuriyan J: **Crystal structure of the nucleotide exchange factor GrpE bound to the ATPase domain of the molecular chaperone DnaK**. *Science* 1997, **276**:431–435.

250. Dragovic Z, Broadley SA, Shomura Y, Bracher A, Hartl FU: **Molecular chaperones of the Hsp110 family act as nucleotide exchange factors of Hsp70s**. *EMBO J* 2006, **25**:2519–2528.

251. Sargeant TJ, Marti M, Caler E, Carlton JM, Simpson K, Speed TP, Cowman AF: **Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites**. *Genome Biol* 2006, **7**:R12.

252. Sharma YD: **Structure and possible function of heat-shock proteins in Falciparum malaria**. *Comp Biochem Physiol - B Biochem Mol Biol* 1992, **102**:437–444.

253. Shonhai A, Boshoff A, Blatch GL: **Plasmodium falciparum heat shock protein 70 is able to suppress the thermosensitivity of an Escherichia coli DnaK mutant strain**. *Mol Genet Genomics* 2005, **274**:70–78.

254. Bell SL, Chiang AN, Brodsky JL: **Expression of a malarial Hsp70 improves defects in chaperone-dependent activities in ssa1 mutant yeast**. *PLoS One* 2011, **6**:e20047.

255. Misra G, Ramachandran R: **Hsp70-1 from Plasmodium falciparum: Protein stability, domain analysis and chaperone activity**. *Biophys Chem* 2009, **142**:55–64.

256. Ramya TNC, Karmodiya K, Surolia A, Surolia N: **15-Deoxyspergualin primarily targets the trafficking of apicoplast proteins in Plasmodium falciparum**. *J Biol Chem* 2007, **282**:6388–6397.

257. Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, Roos DS, Cowman AF, McFadden GI: **Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum**. *Science* 2003, **299**:705–708.

258. Sayeed SK, Shah V, Chaubey S, Singh M, Alampalli S V, Tatu US: **Identification of heat shock factor binding protein in Plasmodium falciparum**. *Malar J* 2014, **13**:118.

259. Acharya P, Pallavi R, Chandran S, Chakravarti H, Middha S, Acharya J, Kochar S, Kochar D, Subudhi A, Boopathi AP, Garg S, Das A, Tatu U: **A glimpse into the clinical proteome of human malaria parasites Plasmodium falciparum and Plasmodium vivax**. *Proteomics - Clin Appl* 2009, **3**:1314–1325.

260. Külzer S, Charnaud S, Dagan T, Riedel J, Mandal P, Pesce ER, Blatch GL, Crabb BS, Gilson PR, Przyborski JM: **Plasmodium falciparum-encoded exported hsp70/hsp40 chaperone/co-chaperone complexes within the host erythrocyte**. *Cell Microbiol* 2012, **14**:1784–1795.

261. Grover M, Chaubey S, Ranade S, Tatu U: **Identification of an exported heat shock protein 70 in Plasmodium falciparum**. *Parasite* 2013, **20**:2.

262. Kappes B, Suetterlin BW, Hofer-Warbinek R, Humar R, Franklin RM: **Two major phosphoproteins of Plasmodium falciparum are heat shock proteins**. *Mol Biochem Parasitol* 1993, **59**:83–94.

263. Peterson MG, Crewther PE, Thompson JK, Corcoran LM, Coppel RL, Brown G V, Anders RF, Kemp DJ: **A second antigenic heat shock protein of Plasmodium falciparum**. *DNA* 1988, **7**:71–78.

264. Šlapeta J, Keithly JS: **Cryptosporidium parvum mitochondrial-type HSP70 targets homologous and heterologous mitochondria**. *Eukaryot Cell* 2004, **3**:483–494.

265. Nyalwidhe J, Lingelbach K: **Proteases and chaperones are the most abundant proteins in the parasitophorous vacuole of Plasmodium falciparum-infected erythrocytes**. *Proteomics* 2006, **6**:1563–1573.

266. Easton DP, Kaneko Y, Subjeck JR: **The Hsp110 and Grp1 70 stress proteins: Newly recognized relatives of the Hsp70s**. *Cell Stress Chaperones* 2000, **5**:276–290.

267. Nathan DF, Vos MH, Lindquist S: **In vivo functions of the Saccharomyces cerevisiae Hsp90 chaperone**. *Proc Natl Acad Sci U S A* 1997, **94**:12949–12956.

268. Li J, Buchner J: **Structure, function and regulation of the hsp90 machinery**. *Biomed J* 2012, **36**:106–17.

269. Ali MMU, Roe SM, Vaughan CK, Meyer P, Panaretou B, Piper PW, Prodromou C, Pearl LH: **Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex**. *Nature* 2006, **440**:1013–1017.

270. Makhnevych T, Houry WA: **The role of Hsp90 in protein complex assembly**. *Biochim Biophys Acta - Mol Cell Res* 2012, **1823**:674–682.

271. Weikl T, Muschler P, Richter K, Veit T, Reinstein J, Buchner J: **C-terminal regions of Hsp90 are important for trapping the nucleotide during the ATPase cycle**. *J Mol Biol* 2000, **303**:583–592.

272. Pearl LH, Prodromou C: **Structure and in vivo function of Hsp90**. *Curr Opin Struct Biol* 2000, **10**:46–51.

273. Prodromou C, Panaretou B, Chohan S, Siligardi G, O'Brien R, Ladbury JE, Roe SM, Piper PW, Pearl LH: **The ATPase cycle of Hsp90 drives a molecular "clamp" via transient dimerization of the N-terminal domains**. *EMBO J* 2000, **19**:4383–4392.

274. Meyer P, Prodromou C, Hu B, Vaughan C, Roe SM, Panaretou B, Piper PW, Pearl LH: **Structural and functional analysis of the middle segment of Hsp90: Implications for ATP hydrolysis and client protein and cochaperone interactions**. *Mol Cell* 2003, **11**:647–658.

275. Prodromou C, Roe SM, O'Brien R, Ladbury JE, Piper PW, Pearl LH: **Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone**. *Cell* 1997, **90**:65–75.

276. Louvion JF, Warth R, Picard D: **Two eukaryote-specific regions of Hsp82 are dispensable for its viability and signal transduction functions in yeast**. *Proc Natl Acad Sci U S A* 1996, **93**:13937–13942.

277. Nemoto T, Ohara-Nemoto Y, Ota M, Takagi T, Yokoyama K: **Mechanism of dimer formation of the 90-kDa heat-shock protein**. *Eur J Biochem* 1995, **233**:1–8.

278. Li J, Soroka J, Buchner J: **The Hsp90 chaperone machinery: Conformational dynamics and regulation by co-chaperones**. *Biochim Biophys Acta - Mol Cell Res* 2012:624–635.

279. Richter K, Walter S, Buchner J: **The co-chaperone Sba1 connects the ATPase reaction of Hsp90 to the progression of the chaperone cycle**. *J Mol Biol* 2004, **342**:1403–1413.

280. Panaretou B, Siligardi G, Meyer P, Maloney A, Sullivan JK, Singh S, Millson SH, Clarke PA, Naaby-Hansen S, Stein R, Cramer R, Mollapour M, Workman P, Piper PW, Pearl LH, Prodromou C: **Activation of the ATPase activity of Hsp90 by the stress-regulated cochaperone Aha1**. *Mol Cell* 2002, **10**:1307–1318.

281. Felts SJ, Toft DO: **p23, a simple protein with complex activities**. *Cell Stress Chaperones* 2003, **8**:108–113.

282. Wegele H, Wandinger SK, Schmid AB, Reinstein J, Buchner J: **Substrate transfer from the chaperone Hsp70 to Hsp90**. *J Mol Biol* 2006, **356**:802–811.

283. Pavithra SR, Kumar R, Tatu U: **Systems analysis of chaperone networks in the malarial parasite Plasmodium falciparum**. *PLoS Comput Biol* 2007, **3**:1701–1715.

284. Pavithra SR, Banumathy G, Joy O, Singh V, Tatu U: **Recurrent fever promotes Plasmodium falciparum development in human erythrocytes**. *J Biol Chem* 2004, **279**:46692–46699.

285. Kampinga HH, Craig EA: **The HSP70 chaperone machinery: J proteins as drivers of functional specificity**. *Nat Rev Mol Cell Biol* 2010, **11**:579–592.

286. Njunge JM, Ludewig MH, Boshoff A, Pesce E-R, Blatch GL: **Hsp70s and J proteins of Plasmodium parasites infecting rodents and primates: Structure, function, clinical relevance, and drug targets**. *Curr Pharm Des* 2013, **19**:387–403.

287. Laufen T, Mayer MP, Beisel C, Klostermeier D, Mogk A, Reinstein J, Bukau B: **Mechanism of regulation of hsp70 chaperones by DnaJ cochaperones**. *Proc Natl Acad Sci U S A* 1999, **96**:5452–5457.

288. Cheetham ME, Caplan AJ: **Structure, function and evolution of DnaJ: Conservation and adaptation of chaperone function**. *Cell Stress Chaperones* 1998, **3**:28–36.

289. Tsai J, Douglas MG: **A conserved HPD sequence of the J-domain is necessary for YDJ1 stimulation of Hsp70 ATPase activity at a site distinct from substrate binding**. *J Biol Chem* 1996, **271**:9347–9354.

290. Suh W-C, Burkholder WF, Lu CZ, Zhao X, Gottesman ME, A GC: **Interaction of the Hsp70 molecular chaperone, DnaK, with its cochaperone DnaJ**. *Proc Natl Acad Sci U S A* 1998, **95**:15223–15228.

291. Jiang J, Maes EG, Taylor AB, Wang L, Hinck AP, Lafer EM, Sousa R: **Structural Basis of J Cochaperone Binding and Regulation of Hsp70**. *Mol Cell* 2007, **28**:422–433.

292. Genevaux P, Schwager F, Georgopoulos C, Kelley WL: **Scanning mutagenesis identifies amino acid residues essential for the in vivo activity of the Escherichia coli DnaJ (Hsp40) J-domain**. *Genetics* 2002, **162**:1045–1053.

293. Botha M, Pesce ER, Blatch GL: **The Hsp40 proteins of Plasmodium falciparum and other apicomplexa: Regulating chaperone power in the parasite and the host**. *Int J Biochem Cell Biol* 2007, **39**:1781–1803.

294. Watanabe J: **Cloning and characterization of heat shock protein DnaJ homologues from Plasmodium falciparum and comparison with ring infected erythrocyte surface antigen**. *Mol Biochem Parasitol* 1997, **88**:253–258.

295. Kumar A, Tanveer A, Biswas S, Ram EVSR, Gupta A, Kumar B, Habib S: **Nuclear-encoded DnaJ homologue of Plasmodium falciparum interacts with replication ori of the apicoplast genome**. *Mol Microbiol* 2010, **75**:942–956.

296. Botha M, Chiang AN, Needham PG, Stephens LL, Hoppe HC, Külzer S, Przyborski JM, Lingelbach K, Wipf P, Brodsky JL, Shonhai A, Blatch GL: **Plasmodium falciparum encodes a single cytosolic type I Hsp40 that functionally interacts with Hsp70 and is upregulated by heat shock.** *Cell Stress Chaperones* 2011, **16**:389–401.

297. Pesce ER, Acharya P, Tatu U, Nicoll WS, Shonhai A, Hoppe HC, Blatch GL: **The Plasmodium falciparum heat shock protein 40, Pfj4, associates with heat shock protein 70 and shows similar heat induction and localisation patterns**. *Int J Biochem Cell Biol* 2008, **40**:2914–2926.

298. Schmid AB, Lagleder S, Gräwert MA, Röhl A, Hagn F, Wandinger SK, Cox MB, Demmer O, Richter K, Groll M, Kessler H, Buchner J: **The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop**. *EMBO J* 2012, **31**:1506–1517.

299. Southworth DR, Agard DA: **Client-Loading Conformation of the Hsp90 Molecular Chaperone Revealed in the Cryo-EM Structure of the Human Hsp90:Hop Complex**. *Mol Cell* 2011, **42**:771–781.

300. McLaughlin SH, Smith HW, Jackson SE: **Stimulation of the weak ATPase activity of human hsp90 by a client protein**. *J Mol Biol* 2002, **315**:787–798.

301. Banumathy G, Singh V, Tatu U: **Host chaperones are recruited in membrane-bound complexes by Plasmodium falciparum**. *J Biol Chem* 2002, **277**:3902–3912.

302. Gitau GW, Mandal P, Blatch GL, Przyborski J, Shonhai A: **Characterisation of the Plasmodium falciparum Hsp70-Hsp90 organising protein (PfHop)**. *Cell Stress Chaperones* 2012, **17**:191–202.

303. Zeytuni N, Zarivach R: **Structural and functional discussion of the tetra-trico-peptide repeat, a protein interaction module**. *Structure* 2012:397–405.

304. Kajander T, Sachs JN, Goldman A, Regan L: **Electrostatic interactions of Hsp-organizing protein tetratricopeptide domains with Hsp70 and Hsp90: Computational analysis and protein engineering**. *J Biol Chem* 2009, **284**:25364–25374.

305. Flom G, Behal RH, Rosen L, Cole DG, Johnson JL: **Definition of the minimal fragments of Sti1 required for dimerization, interaction with Hsp70 and Hsp90 and in vivo functions**. *Biochem J* 2007, **404**:159–167.

306. Scheufler C, Brinker A, Bourenkov G, Pegoraro S, Moroder L, Bartunik H, Hartl FU, Moarefi I: **Structure of TPR domain-peptide complexes: Critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine**. *Cell* 2000, **101**:199–210.

307. Song Y, Masison DC: **Independent regulation of Hsp70 and Hsp90 chaperones by Hsp70/Hsp90-organizing protein Sti1 (Hop1)**. *J Biol Chem* 2005, **280**:34178–34185.

308. Onuoha SC, Coulstock ET, Grossmann JG, Jackson SE: **Structural Studies on the Co-chaperone Hop and Its Complexes with Hsp90**. *J Mol Biol* 2008, **379**:732–744.

309. Richter K, Muschler P, Hainzl O, Reinstein J, Buchner J: **Sti1 is a non-competitive inhibitor of the Hsp90 ATPase. Binding prevents the N-terminal dimerization reaction during the atpase cycle.** *J Biol Chem* 2003, **278**:10328–33.

310. Whitesell L, Mimnaugh EG, De Costa B, Myers CE, Neckers LM: **Inhibition of heat shock protein HSP90-pp60v-src heteroprotein complex formation by benzoquinone ansamycins: Essential role for stress proteins in oncogenic transformation**. *Proc Natl Acad Sci U S A* 1994, **91**:8324–8328.

311. Schneider C, Sepp-Lorenzino L, Nimmesgern E, Ouerfelli O, Danishefsky S, Rosen N, Hartl FU: **Pharmacologic shifting of a balance between protein refolding and degradation mediated by Hsp90**. *Proc Natl Acad Sci U S A* 1996, **93**:14536–14541.

312. Angel SO, Matrajt M, Echeverria PC: **A review of recent patents on the protozoan parasite HSP90 as a drug target**. *Recent Pat Biotechnol* 2013, **7**:2–8.

313. Jhaveri K, Taldone T, Modi S, Chiosis G: **Advances in the clinical development of heat shock protein 90 (Hsp90) inhibitors in cancers**. *Biochim Biophys Acta - Mol Cell Res* 2012, **1823**:742–755.

314. Jego G, Hazoumé A, Seigneuric R, Garrido C: **Targeting heat shock proteins in cancer**. *Cancer Lett* 2013, **332**:275–285.

315. Shahinas D, Liang M, Datti A, Pillai DR: **A repurposing strategy identifies novel synergistic inhibitors of plasmodium falciparum heat shock protein 90**. *J Med Chem* 2010, **53**:3552–3557.

316. Taldone T, Ochiana SO, Patel PD, Chiosis G: **Selective targeting of the stress chaperome as a therapeutic strategy**. *Trends Pharmacol Sci* 2014, **35**:592–603.

317. Matts RL, Brandt GEL, Lu Y, Dixit A, Mollapour M, Wang S, Donnelly AC, Neckers L, Verkhivker G, Blagg BSJ: **A systematic protocol for the characterization of Hsp90 modulators**. *Bioorganic Med Chem* 2011, **19**:684–692.

318. Cortajarena AL, Yi F, Regan L: **Designed TPR modules as novel anticancer agents**. *ACS Chem Biol* 2008, **3**:161–166.

319. Pesce E-R, Cockburn IL, Goble JL, Stephens LL, Blatch GL: **Malaria heat shock proteins: drug targets that chaperone other drug targets**. *Infect Disord Drug Targets* 2010, **10**:147–157.

320. Wright CM, Chovatiya RJ, Jameson NE, Turner DM, Zhu G, Werner S, Huryn DM, Pipas JM, Day BW, Wipf P, Brodsky JL: **Pyrimidinone-peptoid hybrid molecules with distinct effects on molecular chaperone function and cell proliferation**. *Bioorganic Med Chem* 2008, **16**:3291–3301.

321. Murphy ME: **The HSP70 family and cancer**. *Carcinogenesis* 2013, **34**:1181–1188.

322. Cockburn IL: **Modulation of Plasmodium falciparum chaperones PfHsp70-1 and PfHsp70-x by small molecules**. Rhodes University; 2012.

323. Hatherley R: **In Silico Characterisation of the Four Canonical Plasmodium falciparum 70 kDa Heat Shock Proteins**. Rhodes University; 2012.

324. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller J a, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ, Treatman C, Wang H: **PlasmoDB: A functional genomic database for malaria parasites.** *Nucleic Acids Res* 2009, **37**(Database issue):D539–43.

325. **NCBI Protein Database** [http://www.ncbi.nlm.nih.gov/protein/]

326. Söding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W244–W248.

327. Pei J, Kim BH, Grishin N V: **PROMALS3D: A tool for multiple protein sequence and structure alignments**. *Nucleic Acids Res* 2008, **36**:2295–2300.

328. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints**. *J Mol Biol* 1993, **234**:779–815.

329. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM: **MetaMQAP: A meta-server for the quality assessment of protein models**. *BMC Bioinformatics* 2008, **9**:403.

330. Bowie JU, Lüthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure**. *Science* 1991, **253**:164–170.

331. Sippl MJ: **Recognition of errors in three-dimensional structures of proteins**. *Proteins Struct Funct Genet* 1993, **17**:355–362.

332. Wiederstein M, Sippl MJ: **ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W407–W410.

333. Jiang J, Prasad K, Lafer EM, Sousa R: **Structural basis of interdomain communication in the Hsc70 chaperone**. *Mol Cell* 2005, **20**:513–524.

334. Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ERP: **Solution conformation of wild-type E. coli Hsp70 (DnaK) chaperone complexed with ADP and substrate**. *Proc Natl Acad Sci U S A* 2009, **106**:8471–8476.

335. Tastan Bishop Ö, Kroon M: **Study of protein complexes via homology modeling, applied to cysteine proteases and their protein inhibitors**. *J Mol Model* 2011, **17**:3163–3172.

336. Schrödinger LLC: **The PyMOL Molecular Graphics System**. .

337. Chaudhury S, Lyskov S, Gray JJ: **PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta**. *Bioinformatics* 2010:689–691.

338. De Vries SJ, van Dijk M, Bonvin AMJJ: **The HADDOCK web server for data-driven biomolecular docking**. *Nat Protoc* 2010, **5**:883–897.

339. Tina KG, Bhadra R, Srinivasan N: **PIC: Protein Interactions Calculator**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W473–W476.

340. Kim DE, Chivian D, Baker D: **Protein structure prediction and analysis using the Robetta server**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W526–W531.

341. Pasini E, Kirkegaard M, Mortensen P: **In-depth analysis of the membrane and cytosolic proteome of red blood cells**. *Blood* 2006, **108**:791–801.

342. Van Gestel RA, van Solinge WW, van der Toorn HWP, Rijksen G, Heck AJR, van Wijk R, Slijper M: **Quantitative erythrocyte membrane proteome analysis with Blue-Native/SDS PAGE**. *J Proteomics* 2010, **73**:456–465.

343. Berjanskii M V, Riley MI, Xie A, Semenchenko V, Folk WR, Van Doren SR: **NMR structure of the N-terminal J domain of murine polyomavirus T antigens: Implications for DnaJ-like domains and for mutations of T antigens**. *J Biol Chem* 2000, **275**:36094–36103.

344. Hennessy F, Nicoll WS, Zimmermann R, Cheetham ME, Blatch GL: **Not all J domains are created equal: Implications for the specificity of Hsp40-Hsp70 interactions**. *Protein Sci* 2005, **14**:1697–1709.

345. Nicoll WS, Botha M, McNamara C, Schlange M, Pesce E-R, Boshoff A, Ludewig MH, Zimmermann R, Cheetham ME, Chapple JP, Blatch GL: **Cytosolic and ER J-domains of mammalian and parasitic origin can functionally interact with DnaK**. *Int J Biochem Cell Biol* 2007, **39**:736–751.

346. Ahmad A, Bhattacharya A, Mcdonald RA, Cordes M, Ellington B, Bertelsen EB, Zuiderweg ERP: **Heat shock protein 70 kDa chaperone/DnaJ cochaperone complex employs an unusual dynamic interface**. *Proc Natl Acad Sci U S A* 2011, **108**:18966–18971.

347. Johnson JL, Halas A, Flom G: **Nucleotide-dependent interaction of Saccharomyces cerevisiae Hsp90 with the cochaperone proteins Sti1, Cpr6, and Sba1**. *Mol Cell Biol* 2007, **27**:768–776.

348. Atilgan C, Atilgan AR: **Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein**. *PLoS Comput Biol* 2009, **5**:e1000544.

349. Epstein CJ, Goldberger RF, Anfinsen CB: **The Genetic Control of Tertiary Protein Structure: Studies With Model Systems**. *Cold Spring Harb Symp Quant Biol* 1963, **28**:439–449.

350. Tastan Bishop A Özlem, de Beer TAP, Joubert F: **Protein homology modelling and its use in South Africa**. *S Afr J Sci* 2008, **104**:2–6.

351. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235–42.

352. **RCSB Protein Data Bank** [http://www.rcsb.org/pdb]

353. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *J Mol Biol* 1990, **215**:403–410.

354. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389–402.

355. Green JR, Korenberg MJ, Aboul-Magd MO: **PCI-SS: MISO dynamic nonlinear protein secondary structure prediction.** *BMC Bioinformatics* 2009, **10**:222.

356. Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6**:361–365.

357. Söding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics* 2005, **21**:951–60.

358. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server.** *Bioinformatics* 2000, **16**:404–405.

359. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22**:2577–2637.

360. **wwPDB: Validation Reports** [http://wwpdb.org/validation]

361. Dorn M, Barbachan M, Buriol LS, Lamb LC: **Three-dimensional protein structure prediction : Methods and computational strategies**. *Comput Biol Chem* 2014, **53**:251–276.

362. Dunbrack RL: **Sequence comparison and protein structure prediction**. *Curr Opin Struct Biol* 2006, **16**:374–384.

363. Sali A, Overington JP, Johnson MS, Blundell TL: **From comparisons of protein sequences and structures to protein modelling and design.** *Trends Biochem Sci* 1990, **15**:235–240.

364. Westhead DR, Thornton JM: **Protein structure prediction.** *Curr Opin Biotechnol* 1998, **9**:383–389.

365. Venclovas Č, Zemla A, Fidelis K, Moult J: **Comparison of performance in successive CASP experiments**. *Proteins Struct Funct Genet* 2001, **5**:163–170.

366. Kopp J, Bordoli L, Battey JND, Kiefer F, Schwede T: **Assessment of CASP7 predictions for template-based modeling targets**. In *Proteins Struct Funct Genet*. *Volume 69*; 2007:38–56.

367. Kryshtafovych A, Fidelis K, Moult J: **CASPB results in context of previous experiments**. *Proteins Struct Funct Bioinforma* 2009, **77**:217–228.

368. Moult J: **Predicting protein three-dimensional structure**. *Curr Opin Biotechnol* 1999, **10**:583–588.

369. Alwyn Jones T, Kleywegt GJ: **CASP3 comparative modeling evaluation.** *Proteins* 1999, **7**:30–46.

370. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Res* 2003, **31**:3497–3500.

371. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C: **3DCoffee: Combining protein sequences and structures within multiple sequence alignments**. *J Mol Biol* 2004, **340**:385–395.

372. Poirot O, Suhre K, Abergel C, O'Toole E, Notredame C: **3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W37–40.

373. Sali A: **Modelling mutations and homologous proteins**. *Curr Opin Biotechnol* 1995, **6**:437–451.

374. Wallner B, Elofsson A: **All are not equal: a benchmark of different homology modeling programs.** *Protein Sci* 2005, **14**:1315–27.

375. Schwede T, Guex N, Peitsch MC, Schwede T: **SWISS-MODEL: an automated protein homology-modeling server**. *Nucleic Acids Res* 2003, **31**:3381–3385.

376. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh IYY, Alexov E, Honig B: **Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling**. *Proteins Struct Funct Genet* 2003, **53**:430–435.

377. Levitt M: **Accurate modeling of protein conformation by automatic segment matching.** *J Mol Biol* 1992, **226**:507–533.

378. Decanniere K, Muyldermans S, Wyns L: **Canonical antigen-binding loop structures in immunoglobulins: more structures, more canonical classes?** *J Mol Biol* 2000, **300**:83–91.

379. Almonte AG, Sweatt JD: **Serine proteases, serine protease inhibitors, and protease-activated receptors: roles in synaptic function and behavior**. *Brain Res* 2011, **1407**:107–122.

380. Lepšík M, Field MJ: **Binding of calcium and other metal ions to the EF-hand loops of calmodulin studied by quantum chemical calculations and molecular dynamics simulations**. *J Phys Chem B* 2007, **111**:10012–10022.

381. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B: **Loop modeling: Sampling, filtering, and scoring**. *Proteins Struct Funct Genet* 2008, **70**:834–843.

382. Du P, Andrec M, Levy RM: **Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update.** *Protein Eng* 2003, **16**:407–414.

383. Sali A: **MODELLER: A Program for Protein Structure Modeling Release 9.14, r10167**. 2014:271.

384. Fiser A, Do RKG, Sali A: **Modeling of loops in protein structures.** *Protein Sci* 2000, **9**:1753–73.

385. Quan L, Lü Q, Li H, Xia X, Wu H: **Improved packing of protein side chains with parallel ant colonies**. *BMC Bioinformatics* 2014, **15**:S5.

386. Xiang Z, Honig B: **Extending the accuracy limits of prediction for side-chain conformations.** *J Mol Biol* 2001, **311**:421–30.

387. Ponder JW, Richards FM: **Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.** *J Mol Biol* 1987, **193**:775–91.

388. Krivov GG, Shapovalov M V., Dunbrack Jr RL: **Improved prediction of protein side-chain conformations with SCWRL4**. *Proteins Struct Funct Bioinforma* 2009, **77**:778–795.

389. Levitt M, Lifson S: **Refinement of protein conformations using a macromolecular energy minimization procedure.** *J Mol Biol* 1969, **46**:269–279.

390. Xiang Z: **Advances in homology protein structure modeling.** *Curr Protein Pept Sci* 2006, **7**:217–27.

391. Ginalski K, Rychlewski L: **Protein Structure Prediction of CASP5 Comparative Modelling and Fold Recognition Targets Using Consensus Alignment Approach and 3D Assessment**. *Proteins Struct Funct Genet* 2003, **53**:410–417.

392. Canutescu AA, Shelenkov AA, Dunbrack RL: **A graph-theory algorithm for rapid protein side-chain prediction.** *Protein Sci* 2003, **12**:2001–2014.

393. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673–4680.

394. Myers EW, Miller W: **Optimal alignments in linear space.** *Comput Appl Biosci* 1988, **4**:1–13.

395. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**:406–25.

396. Larkin M a., Blackshields G, Brown NP, Chenna R, Mcgettigan P a., McWilliam H, Valentin F, Wallace IM, Wilm a., Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**:2947–2948.

397. 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* 2002, **30**:3059–66.

398. Miyata T, Miyazawa S, Yasunaga T: **Two types of amino acid substitutions in protein evolution**. *J Mol Evol* 1979, **12**:219–236.

399. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program**. *Brief Bioinform* 2008, **9**:286–298.

400. Edgar RC: **MUSCLE: A multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.

401. Edgar RC: **MUSCLE: Multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792–1797.

402. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205–217.

403. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48**:443–453.

404. Huang X, Miller W: **A time-efficient, linear-space local similarity algorithm**. *Adv Appl Math* 1991, **12**:337–357.

405. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C: **Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee.** *Nucleic Acids Res* 2006, **34**(Web Server issue):W604–8.

406. Laskowski R a., MacArthur MW, Moss DS, Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures**. *J Appl Crystallogr* 1993, **26**:283–291.

407. Doreleijers JF, Da Silva AWS, Krieger E, Nabuurs SB, Spronk C a EM, Stevens TJ, Vranken WF, Vriend G, Vuister GW: **CING: An integrated residue-based structure validation program suite**. *J Biomol NMR* 2012, **54**:267–283.

408. Elengoe A, Naser MA, Hamdan S: **Modeling and docking studies on novel mutants (K71L and T204V) of the ATPase domain of human heat shock 70 kDa protein 1**. *Int J Mol Sci* 2014, **15**:6797–814.

409. Kherraz K, Kherraz K, Kameli A: **Homology modeling of Ferredoxin-nitrite reductase from Arabidopsis thaliana.** *Bioinformation* 2011, **6**:115–119.

410. Lüthy R, Bowie JU, Eisenberg D: **Assessment of protein models with three-dimensional profiles.** *Nature* 1992, **356**:83–85.

411. Sippl MJ: **Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.** *J Mol Biol* 1990.

412. Melo F, Feytmans E: **Assessing protein structures with a non-local atomic interaction energy.** *J Mol Biol* 1998, **277**:1141–1152.

413. Krishnamoorthy B, Tropsha A: **Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations.** *Bioinformatics* 2003, **19**:1540–8.

414. Lin K, May ACW, Taylor WR: **Threading using neural nEtwork (TUNE): the measure of protein sequence-structure compatibility.** *Bioinformatics* 2002, **18**:1350–1357.

415. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A: **Protein fragment reconstruction using various modeling techniques.** *J Comput Aided Mol Des* 2003, **17**:725–38.

416. Wallner B, Elofsson A: **Identification of correct regions in protein models using structural, alignment, and consensus information.** *Protein Sci* 2006, **15**:900–913.

417. Shen M-Y, Sali A: **Statistical potential for assessment and prediction of protein structures.** *Protein Sci* 2006, **15**:2507–24.

418. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T: **SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information.** *Nucleic Acids Res* 2014, **42**(Web Server issue):W252–8.

419. Benkert P, Tosatto SCE, Schomburg D: **QMEAN: A comprehensive scoring function for model quality assessment.** *Proteins* 2008, **71**:261–277.

420. Zhang Y: **I-TASSER server for protein 3D structure prediction.** *BMC Bioinformatics* 2008, **9**:40.

421. Huang YJ, Mao B, Aramini JM, Montelione GT: **Assessment of template-based protein structure predictions in CASP10**. *Proteins Struct Funct Bioinforma* 2014, **82**:43–56.

422. Krieger E, Koraimann G, Vriend G: **Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field.** *Proteins* 2002, **47**:393–402.

423. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K: **Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8**. *Proteins Struct Funct Bioinforma* 2009, **77**:114–122.