High-throughput prioritization of putative drug targets in the malaria parasite, *Plasmodium falciparum*

by
Misha le Grange
Submitted in partial fulfillment of the requirement for the degree
Magister Scientiae Bioinformatics
in the Faculty of Natural and Agricultural Science
Bioinformatics and Computational Biology Unit
Department of Biochemistry
University of Pretoria
Pretoria

February 11, 2015

Declaration

I, Misha le Grange, declare that the thesis/dissertation which I hereby submit for the degree MSc Bioinformatics at the University of Pretoria is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: ....

Date: ....

Acknowledgements

- I would like to thank Jesus Christ for helping me through all my studies and giving me the strength to keep moving forward.
- I would also like to thank the NRF for funding this project, and providing the opportunity to participate in conferences.
- A special thanks to my parents for always believing in me and enduring it whilst studying
- Also, thanks to my fellow students, especially to Nanette Coetzer for all her support and help with the statistics and figuring out R; and to Oliver Bezuidt.
- A big thank you for Jeanre Smit for all the help with Taverna and the setting up of the pipelines.
- And last but not least, thank you to Fourie Joubert, my supervisors for the assistance and guidance with the project.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	:	Two Dimensional
3D	:	Three Dimensional
AC	:	Uniprot Accession Number
ACK1	:	Activated CDC42 Kinase 1
ACT	:	Artemisini-based Combination Therapy
BLAST	:	Basic Local Alignment Search Tool
CADD	:	Computer Aided Drug Design
COMT	:	Cathol-O-Methyl Transferase
CTD	:	Carboxyl-Terminal Domain
DOPE	:	Discrete Optimized Protein Energy
EC	:	Enzyme Commission
EP	:	Evolutionary Patterning
ER	:	Endoplasmic Reticulum
ERAD	:	Endoplasmic Reticulum-Associated Degradation
ET	:	Evolutionary Tracing
FDA	:	Food and Drug Administration
GO	:	Gene Ontology
GPCR	:	G Protein-Coupled Receptors
GSK	:	Glaxo Smith Kline
GUI	:	Graphic User Interface
HTS	:	High Throughput Screening
InhA	:	Enoyal-acyl carrier protein reductase
LBDD	:	Ligand Based Drug Design

MSA	Multiple Sequence Alignments
NR	Nuclear Receptors
PDB	Protein Databank
PDTD	Potential Drug Target Database
PfCRT	*P. falciparum* Chloroquine Resistance Transporter
PfMDR	*P. falciparum* Multi-Drug Resistance
PI3K	Phosphatidyl inositol 3' kinases
PK	Protein Kinase
QSAR	Quantitative Structure-Activity Relationships
RBC	Red Blood Cells
RP	RNA Polymerases
SAH	S-adenosyl-l-homocysteine
SAHH	S-adenosyl-l-homocysteine hydrolase
SAM	S-adenosylmethionine
SAR	Structure-activity Relationship
SBDD	Structure Based Drug Design
SOAP	Simple Object Access Protocol
SP	Sulfodoxine Pyrimethamine
TarFisDock	Target Fishing Dock
TBP	TATA-Binding Protein
TDR	Tropical Disease Research
TF	Transcription Factors
Ub	Ubiquitin
UBC	Catalytic Core Domain
UBE2N	Ub-conjugating enzyme E2 N
XML	eXtensible Markup Language

Preface

Drug resistance to almost all known antimalarials is widespread and is rapidly increasing. This resistance is due to the over and misuse of these antimalarials, thus new antimalarial drugs are necessary to help in the prevention and cure of this widespread disease. Continuous in-depth studies are being done on a handful of putative targets for future exploitation and use, but not many resources are available that focus on performing data mining and target identification on the complete malaria genome, together with relations to chemical compounds.

The DISCOVERY Database is a web-based system, developed for the *in silico* selection of drug target proteins and lead compounds. It is a database filled with malaria information and aspects that might influence the druggability of a malaria parasite protein and guide a scientist in choosing the right ligand for a protein. DISCOVERY can aid in attempting to predict the interaction of ligands with proteins of interest, associating chemical compound with malaria proteins and selective chemical similarity searches. It can be used to mine information on malaria proteins, predict ligands and compare human and mosquito host characteristics.

DISCOVERY2 was developed in Java with NetBeans. The protein sequences for the *Plasmodium* spp. included in DISCOVERY were downloaded from PlasmoDB; the *Homo sapiens* proteins were downloaded from Ensembl and the *Anopheles gambiae* proteins was downloaded from VectorBase.

Even though DISCOVERY is primarily focused on *Plasmodium falciparum* it also contains information for all proteins from *Plasmodium vivax*, *Plasmodium yoelii*, *Plasmodium knowlesi*, *Plasmodium chabaudi* and *Plasmodium berghei* as well for the human vector and mosquito host. Protein information includes sequences and annotations, functional predictions, gene ontology terms, orthology information, structural information, metabolic pathways, predicted putative protein-ligand interactions, druggability predictions and literature links. Chemical compounds are also included.

Recently approaches have illustrated the value of predicting the association of chemical compounds with putative drug targets, especially when the targets of compounds, like the Glaxo Smith Kline dataset with known activity against the parasite may be extrapolated, using protein-ligand interaction databases, like ChemProt. DISCOVERY attempts to use a similar approach in associating chemical compounds with malaria proteins, using sequence homology, and also selective chemical similarity searches.

Chapter 1 of this dissertation is a literature review focusing on the *in silico* identification of potential drug targets. It also mentions a few techniques/approaches with which to accomplish this as well as target databases that can be used to help in the identification process. *Chapter 2* describes the steps taken to run and score the *Plasmodium falciparum* proteins in a high throughput manner through DISCOVERY. *Chapter 3* gives four case studies from DISCOVERY, a protein that had a low weighted score, a protein with a very high weighted score and two proteins with weighted scores in between the other two. And *Chapter 4* concludes by looking at how researchers can use this study as a starting point.

In this dissertation, DISCOVERY2 was used, in conjunction with Taverna pipelines, to study all *Plasmodium falciparum* proteins in a high throughput manner to be able to identify possible drug targets that might be of importance for future drug identification.

Chapter 1

In silico identification of potential drug targets

1.1 Introduction

According to Li and Lai (2007), Imming *et al.* (2006) gave the following definition of a drug target: "*. . . a target to be a molecular structure (chemically definable by at least a molecular mass) that will undergo a specific interaction with chemicals that we call drugs because they are administered to treat or diagnose a disease. The interaction has a connection with the clinical effect(s)*".

1.1.1 Drug Discovery

Widespread resistance to many known antipathogenic drugs is increasing at an alarming rate and is posing a global health threat. Thus, new drug target proteins and lead compounds need to be urgently identified (Joubert *et al.*, 2009; Koseki *et al.*, 2013). The identification of interactions between drug and target proteins plays an important role in the process of genomic drug discovery (Cao *et al.*, 2012).

The cost of research and development in the pharmaceutical industry has been rising steeply and steadily in the last decade, but the amount of time required to bring a new product to market remains around ten to fifteen years (Hasan *et al.*, 2006). Developing analytical methods that facilitate the discovery of some of the general rules for discovering targets for therapeutic agents, and the effects of drug-target interactions, both beneficial and adverse, would be valuable in moving the drug discovery process forward (Ma'ayan *et al.*, 2007).

1.1.2 Drug Target Discovery

Drug discovery begins with the identification (and validation) of a potential target that is then subjected to high throughput screening against a library of drug-like compounds or to rational drug design (Bakheet and Doig, 2009; Florez *et al.*, 2010; Hasan *et al.*, 2006; Li and Lai, 2007; Liu *et al.*, 2010). This can provide the foundation for years of dedicated research in the pharmaceutical industry (Hasan *et al.*, 2006).

Target-based drug discovery has come to virtually eclipse physiology-based approaches toward therapeutic development across the medical spectrum; but it has not necessarily eclipsed pathophysiology-based approaches. Physiology-based drug approaches tended to be low throughput and the practitioners had a tendency to disregard the disease's mechanism, focusing on symptomatic treatments while target-based drug discovery tend to be more compliant to high-throughput assays and their uses aids in rational drug discovery programs (Hilbush *et al.*, 2005).

Presently the most frequent protein targets for which successful drugs have been developed includes 130 protein families which generally includes enzymes, like proteases and kinases; G protein-coupled receptors (GPCR); ion channels and transporters; and nuclear receptors (NR) (Bakheet and Doig, 2009; Bleakley and Yamanishi, 2009; Li and Lai, 2007). Of these GPCRs (23%) and enzymes (50%) represent the most important target classes (Bakheet and Doig, 2009; Bleakley and Yamanishi, 2009). Currently the first step in drug discovery tends to be the identification of the organism's pathways and identifying pathways unique to the organism, compared to its host's pathways. From these potential drug targets can be identified (Bhasme *et al.*, 2013; Ravindranath *et al.*, 2013).

1.1.2.1 Enzymes

Enzymes are a large class of functional proteins that catalyse chemical reactions. Almost all chemical reactions in a biological cell need enzymes as a catalyst for it to function at rates sufficient for life. Identifying drug-enzyme interactions might have a direct application for completing genome annotations, finding enzymes for synthetic chemistry and predicting drug specificity, promiscuity and pharmacology (Cao *et al.*, 2012).

Enzymes, especially protein kinases, are good candidate drug targets because they display global effects through their ability to phosphorylate many targets exhibiting broad changes in cell behaviour (Ma'ayan and He, 2010). Four tRNA synthetases have been proposed as drug targets in *Brugia* and other parasite-causing diseases as it must be essential for the parasite, but are structurally different with respect to human orthologs (Crowther *et al.*, 2010).

New drug targets needed to be discovered for gram-negative bacteria, as drug resistance is a big problem. Advances in Structure-Based Drug Discovery (SBDD) technology, combined with understanding of the factors that influence gram-negative permeability and drug efflux has made it possible to design aspects important for broad-spectrum antibacterial agents. Target selection is central to this process and needs to meet certain criteria. These are:

1. The target active-sites need characteristics that allow for the design of highly potent enzyme inhibitors;
2. The inhibitor target needs to be unique to bacteria, compared to the human host, but at the same time conserved amongst the bacteria class;
3. The inhibitor-binding site needs to be distinct from the sites targeted by existing drugs to avoid cross-resistance with established antibiotic classes;
4. It is desirable to find conserved target pairs from different essential pathways that could be inhibited by a single agent and,
5. The active-site of the target needs to be compatible with inhibitors possessing features necessary for gram-negative penetration and retention.

The ATP binding subunits of the bacterial topoisomerases DNA gyrase (GyrB) and topoisomerase IV (ParE) met these criteria. These topoisomerase complexes are validated drug targets. Further optimization is still needed until these targets are viable drug targets for broad-spectrum gram-negative antibiotics (Tari *et al.*, 2013).

1.1.2.2 G protein-coupled receptors

The GPCR superfamily is the largest known class of cell surface targets. GPCRs might share similar structure by having seven trans membrane domains, it is extremely diverse with capacity to change messages triggered by different stimuli, that might include ligands such as photons, organic odorants, nucleotides, nucleosides, peptides, lipids and even proteins (Cao *et al.*, 2012; Goswami *et al.*, 2013). When a GPCR agonist binds to the extracellular domain of the GPCR, it induces a change in conformation of the receptor leading to coupling and to the activation of one or more G proteins inside the cell.

Thus, GPCRs are membrane embedded proteins responsible for communication between the cell and its environment (Becker *et al.*, 2004; Cao *et al.*, 2012; Pan *et al.*, 2008). As a consequence, many major diseases can involve the malfunction of these receptors, making them among the most important drug targets for pharmacological intervention (Becker *et al.*, 2004), comprising > 50% of pharmacotherapies available on the market. But only a third of these GPCRs have been explored for drug development indicating a future active area of research for the discovery of novel therapeutics. Polymorphisms in GPCRs can affect drug efficacy through altered ligand binding, receptor activation/inactivation, and/or varied signalling cascades (Goswami *et al.*, 2013).

GPCRs regulate the function of ion channels, which play an essential role in the function of neurons by mediating electrical currents and regulation of selective ion concentrations across the cell membrane. Thus, GPCRs are mostly found in the peripheral and central nervous systems where they are located on the plasma membrane of neurons along the nociceptive pathways. GPCRs have been shown to play an important role in the modulation of pain and are one of the most important therapeutic targets in the area of pain management (Pan *et al.*, 2008).

There are seven GPCR receptors that play a role in pain management, these are:

1. Opioid receptors; can bind to drugs like morphine and codeine;
2. Cannabinoid receptors; can bind to marijuana for antagonistic effect;
3. α2-Adrenergic receptors; bind to drugs like: Clonidine, Guanfacine, Guanabenz, Guanoxabenz, Guanethidine, Xylazine, Tizanidine, Methyldopa, Fadolmidine and Dexmedetomidine;
4. Muscarinic acetylcholine receptors; bind to drugs like: Atropine, Benadryl, Benztropine, Biperiden, Ipratropium, Oxitropium, Tiotropium, Glycopyrrolate, Oxybutynin, Tolterodine, Chlorpheniramine, Diphenhydramine, Dimenhydrinate, Orphenadrine, Trihexyphenidyl and Dicyclomine;
5. GABAB receptors; antagonist include Saclofen and Phaclofen;
6. Group II and III metabotropic glutamate receptors; bind to antagonist LY-341, 495 and MGS-0039 and,
7. Somatostatin receptors (Pan *et al.*, 2008).

1.1.2.3 Ion channels

Ion channels are a large superfamily of membrane proteins that pass ions across membranes and are critical to diverse physiological functions in both excitable and non-excitable cells, thus underlie many diseases. Because of this ion channels are an important target class, which has been proven highly “druggable” (Cao *et al.*, 2012).

1.1.2.4 Nuclear receptors

NR are ligand-activated transcription factors that regulate a variety of functions, e.g. homeostasis, reproduction, development and metabolism. Nuclear hormone receptors function as ligand-activated transcription factors, providing a direct link between signalling molecules that control these processes and transcriptional responses.

NR bind small molecules that can easily be modified by drug design, and control functions associated with major diseases. It is an important drug target in terms of potential therapeutic applications and is estimated that 13% of the Food And Drug administration's (FDA) approved drugs have NRs as molecular targets (Cao *et al.*, 2012).

Until 2000, only approximately 500 drug targets had been reported (Gao *et al.*, 2008; Li and Lai, 2007), of these only 120 drug targets are actually marketed and available for the public (Gao *et al.*, 2008). This indicates that the pharmaceutical industry relies on a small pool of drug targets (Li and Lai, 2007). The emergence of molecular medicine and the completion of the human genome and numerous pathogen genomes provide more opportunity to discover unknown target proteins for drugs and suggests that there are 30,000 to 40,000 genes across the species and at least the same number of proteins that are potential targets for drug discovery (Cao *et al.*, 2012; Gao *et al.*, 2008).

Recent advances in genomics have triggered a shift in drug discovery from the paradigm of focusing on strong single-target interaction to more global and comparative analysis of multi-target networks (Liu *et al.*, 2010). In this context; and since experimental investigations of possible drug targets are time-consuming and expensive; it is worthwhile to conduct *in silico* analysis (Crowther *et al.*, 2010; Khoshkholgh-Sima *et al.*, 2011; Li and Lai, 2007; Liu *et al.*, 2010; Subramanian *et al.*, 2006). These analyses can be fast, robust and an efficient method to identify and validate new druggable targets (Liu *et al.*, 2010).

1.1.3 *In silico* discovery

It is time consuming and costly to determine drug-target interactions by experiments alone (Aparoy *et al.*, 2012; Cao *et al.*, 2012), thus drug target identification by *in silico* methods has emerged. This has caused a phenomenal achievement in the field of drug discovery (Bleakley and Yamanishi, 2009) giving an effective alternative to unaffordable high throughput *in vitro* target profiling of compounds as well as to find new therapeutic indications for old drugs (Li *et al.*, 2006; Liu *et al.*, 2010).

Traditional approaches rely on a step-wise synthesis and screening of large number of compounds to identify a potential candidate, thus it is of great practical interest to develop genuinely effective Computer Aided Drug Design (CADD) (*in silico*) methods which can both provide new predictions to experimentalists and provide supporting evidence to experimental results (Aparoy *et al.*, 2012; Bleakley and Yamanishi, 2009). CADD has already been widely used in lead identification and lead optimization stages of drug development against various targets over the years (Aparoy *et al.*, 2012). Compared to traditional drug discovery approaches these methods decrease the time and cost involved in the drug development process (Aparoy *et al.*, 2012).

A variety of approaches have been developed to analyse and predict compound-protein interactions (Bleakley and Yamanishi, 2009). These include structure- and ligand-based methods. Ligand-based methods are based on the calculations of the molecular properties of the compounds while structure-based approaches are based on the study of the interaction between compounds and their target proteins (Jorgensen, 2010; Vilar and Costanzi, 2012).

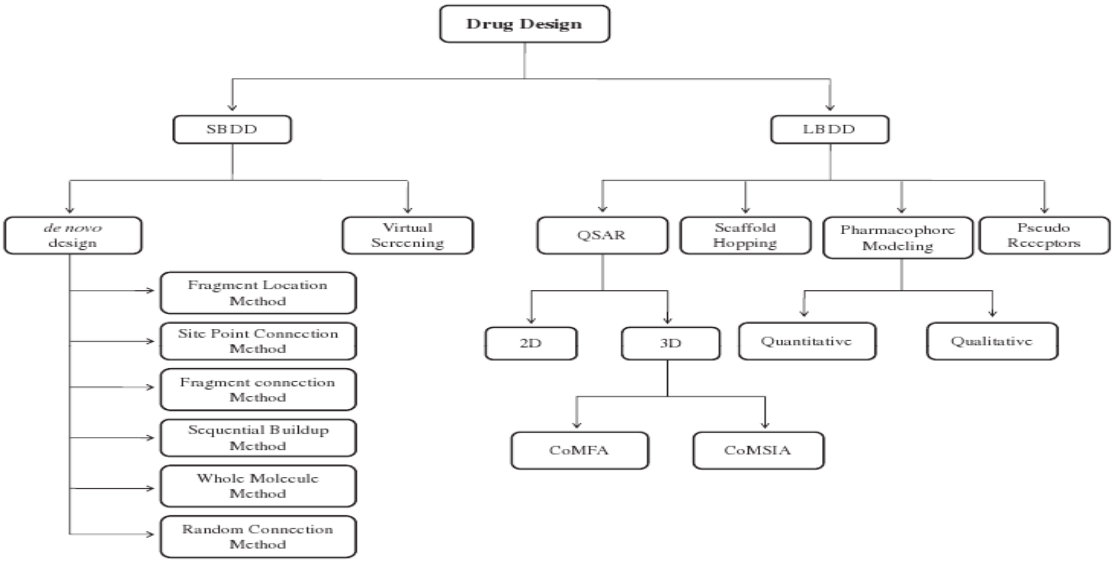


Figure 1.1: Basic principles and types of drug design.
SBDD is the various approaches that use structural information of the drug target while LBDD is the approach where the information for existing ligands of a drug target is used.
From: Aparoy *et al.* (2012).

1.1.3.1 Structure-based drug discovery (SBDD)

SBDD can be done using only the target structure and graphic tools for building ligands in the proposed binding site. But, additional insights provided by evaluating the molecular energetics for the binding process can be central to most current structure-based design activities (Jorgensen, 2010).

For SBDD, structural information is exploited for the development of its inhibitor. The structure can be determined by experimental techniques, like X-ray crystallography or NMR, or it can be determined by predictive techniques such as computational methods like threading and homology modelling (Aparoy *et al.*, 2012).

De novo drug design:

Active site of drug targets, when characterized from a structural point of view, will shed light on its binding features. The information of active site composition and the orientation of the amino acids at the binding site can be used to design ligands specific to that particular target. Computational tools that can analyse protein active site and suggest potential compounds are especially used for *de novo* design methods (Aparoy *et al.*, 2012).

According to Aparoy *et al.* (2012), detailed analysis of CADD can be classified in to six classes (as can be seen in Figure 1.1):

1. Fragment location methods: To determine desirable locations of atoms or small fragments within the active site;
2. Site point connection methods: To determine locations (“site points”) and then place fragments within the active site so that suitable atoms occupy those locations;
3. Fragment connection methods: Fragments are positioned and “linkers” or “scaffolds” are used to connect those fragments and hold them in a desirable orientation;
4. Sequential build-up method: Construct a ligand atom-by-atom, or fragment-by-fragment;

5. Whole molecule methods: Compounds are placed into an active site in various conformations, assessing shape and/or electrostatic complementarity and,
6. Random connection methods: Combing some of the features of fragment connection and sequential build up method, along with bond disconnection strategies and ways to introduce randomness (Aparoy *et al.*, 2012).

An example of the whole molecule method that is frequently used is docking (Aparoy *et al.*, 2012). Docking is a powerful molecular modelling approach that predicts the preferred orientation of the drug molecule to the protein, when bound to each other to form a stable complex by dynamic simulation. A series of ranked drug-target relations can be generated by the size of the energy scores (Cao *et al.*, 2012). An advantage of docking is that it is not reliant on a training set (Vilar and Costanzi, 2012).

Structure-based virtual screening:

This is a commonly-used approach in the lead identification step and is seen as a complementary approach to experimental high throughput screening (HTS) to improve speed and efficiency of the drug discovery and development process. This process involves explicit molecular docking of each ligand to the binding site of the target and scoring. The compounds in the databases screened are ranked with a view to selecting and experimentally testing a small subset for biological activity, considered to be appropriate for a given receptor.

Successful applications have been reported in the field of molecular docking-based virtual screening. Even thought the energy calculations involved are crude, the compounds in the library are readily available, if chosen correctly, making experimental testing easy and false positives tolerable (Aparoy *et al.*, 2012).

1.1.3.2 Ligand-based drug discovery (LBDD)

Ligand-based design does not require the target structure, but rather stems from analysis of structure/activity data for compounds that have been tested in an assay for the biological function of the target. Patterns are looked for in the assay results that might suggest potential modifications of the compounds to yield enhanced activity. An advantage is that no structure is needed, but a disadvantage is that substantial activity data are needed (Jorgensen, 2010).

LBDD approaches predict the drugs interacting with a single given protein based on the chemical structure similarity in a classic structure-activity relationship (SAR) framework (Cao *et al.*, 2012). It includes data mining and Quantitative SAR (QSAR) (Subramanian *et al.*, 2006; Yamanishi *et al.*, 2010).

A limitation of the ligand-based approach includes poor performance when the number of known ligands for a target protein of interest decreases (Yamanishi *et al.*, 2010). Three Dimensional (3D) QSAR and pharmacophore modelling are the most important and widely used tools in LBDD. These can provide predictive models suitable for lead identification and optimization (Aparoy *et al.*, 2012).

QSAR:

An approach, like QSAR, compares a candidate ligand with the known ligands of a target protein to predict it's binding, by using machine learning methods (Yamanishi *et al.*, 2010). QSAR correlate molecular structure with properties like *in vitro* or *in vivo* biological activity by only using ligands with known biological activities (Aparoy *et al.*, 2012; Vilar and Costanzi, 2012). When applied to toxicity data these methods are termed quantitative structure toxicity relationship (QSTR) and when modelled to physiochemical properties it is called quantitative structure property relationship (QSPR).

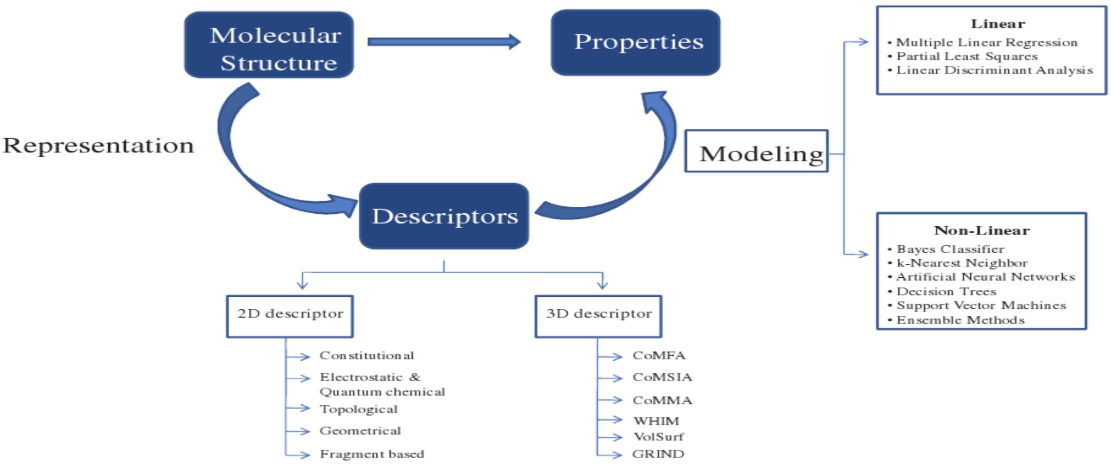


Figure 1.2: General methodology of a QSAR study.
QSAR is the process of studying a series of molecules of different structure and properties and attempting to find empirical relationship between structure and property.
From: Aparoy *et al.* (2012).

This method is equipped with robustness and good ranking ability when applied to the prediction of the activity of closely related analogs (Vilar and Costanzi, 2012). QSAR assumes that the structure of a molecule must contain the features responsible for its physical, chemical and biological activity. It is also defined as a process that quantitatively correlates structural molecular properties with functions for a set of similar compounds. Figure 1.2 depicts a flowchart for QSAR methodology (Aparoy *et al.*, 2012).

A disadvantage of this technique is it is very dependent on a training set; this limits its applicability to the evaluation of diverse compounds (Vilar and Costanzi, 2012).

Pharmacophore modelling strategies:

Pharmacophores refer to the molecular framework that carries the features that are essential for the biological activity of a drug. Peter Gund (1977) defined it as "*a set of structural features in a molecule that are recognized at the receptor site and is responsible for that molecule's biological activity*". It has become a major tool in drug design studies where the 3D structure of the target is known.

An advantage of this approach is rapid screening of millions of compounds for identification of potential candidates. It involves three processes, namely:

1. Finding the features required for a particular biological activity;
2. Determining the molecular conformation required and,
3. Developing a superposition or alignment rule for the series of compounds.

Pharmacophore strategies include sequential computational steps, namely: drug target selection, database preparation, pharmacophore model generation and 3D screening (Aparoy *et al.*, 2012).

Scaffold hopping:

This is a technique that aims to find compounds that are structurally diverse and share a particular biological activity that is, preserving the 3D interaction properties of a scaffold while changing the structural skeleton. If structural diverse compounds are identified this would help in finding new classes of compounds against the target protein. It has been previously described as "*finding isofunctional, but structurally dissimilar molecular entities*" (Aparoy *et al.*, 2012).

Pseudo-receptor modelling:

This is a new concept in CADD that allows the reconstruction of the 3D structure of an unknown target based on the structures of its ligands. It combines present techniques, but extends its possibilities by the generation of an explicit receptor model. This model may then be used for affinity prediction and other receptor-based modelling task. Pseudo-receptors models bridge the gap of ligand- and receptor-based drug design. Pseudo-receptors fall into the class of 3D QSAR methods (Aparoy *et al.*, 2012).
The current state-of-the-art method involves integrative methods that simultaneously take into account things such as target protein sequences, drug chemical structures and the currently known drug target network (Bleakley and Yamanishi, 2009). CADD, based on molecular docking, pharmacophore modelling, 3D quantitative structure activity and molecular dynamics, is a shortcut technique that enables more rapid hit identification than HTS (Koseki *et al.*, 2013).

Systems biology brings the ability to better understand cellular, tissue and organ behaviour at the molecular level. This understanding could lead to better drug design, multi-drug treatment, side-effect prediction, and rapid drug targeting and development as well as biomarker discovery (Ma'ayan *et al.*, 2007). There is a strong inspiration to develop new *in silico* methods capable of detecting potential compound-protein interactions (Yamanishi *et al.*, 2010).

1.2 *In silico* drug discovery for parasites

1.2.1 Criteria

Drug target identification for antiparasitic use is complicated by the fact that the identified drug target must satisfy a variety of criteria to permit progression to the next stage (Hasan *et al.*, 2006), compared to drugs for all indications. One important aspect of a good target is its reliability and resistance to developing resistance over long periods (Iskar *et al.*, 2010).

Certain criteria are influenced by the choice of the target (Perumal *et al.*, 2009). Some criteria might be lacking in the less-studied pathogen, this can be partially overcome by mapping data from homologous genes in well-studied organisms (Crowther *et al.*, 2010). Here are some criteria to be considered when choosing a drug target for a parasite:

1.2.1.1 Is the structure or function known?

Known protein function or structure may make the protein easier to study for drug discovery (Khoshkholgh-Sima *et al.*, 2011), giving it practical advantages for further studies. If the target's structure is available it could aid in rational drug design, providing an advantage in high throughput docking and lead optimization studies (Hasan *et al.*, 2006). Bioinformatics tools have enabled researchers to extract and manipulate the biological information from its sequence, with the goal of understanding the protein function, giving it an insight into biological systems, but thus far, the knowledge of the functions of proteins in their native form has not provided an understanding of cellular behaviour (Florez *et al.*, 2010).

These should not be a primary consideration in identifying a new drug target, as it is mainly of practical consideration (Hasan *et al.*, 2006).

1.2.1.2 Does the protein have a druggable site?

According to *Edfeldt* et al. *(2011)* druggability can be defined as: " *the likelihood of finding orally bioavailable small molecules that bind to a particular target in a disease-modifying way*" (Edfeldt *et al.*, 2011). Structure-based analysis has led to this concept. This is used to describe proteins that possess protein folds that favour interactions with drug-like chemical compounds (Bakheet and Doig, 2009).

By nature, target selection and target validation focuses on molecular biology aspects, but they must also consider the ability of drug-like molecules to bind and alter the biological activity of the target. Most drugs only act on a few target types, thus protein classification can be an indication of druggability (Alvarez-Garcia *et al.*, 2012). It is also important to consider that many proteins are druggable according to their structure, but their binding will not lead to the therapeutic benefit. Therefore druggable proteins are not necessarily drug targets (Bakheet and Doig, 2009).

Proteins that are generally considered druggable are enzymes and receptors, while other protein classes are difficult to use as a target or undruggable. This is a good assumption, as the favoured proteins tend to naturally bind to small organic molecules, which means that drugs can compete in equal terms and achieve high affinity for the binding site. Depending on the type of natural substrate, their difficulty as targets will also vary, e.g. proteins binding *bona fide* small molecules, like class A GPCRs and kinases are more druggable than those binding non-drug-like ligands, like class B and C GPCRs, proteases and other peptides (Alvarez-Garcia *et al.*, 2012).

Analysis of binding site properties can be used to predict protein druggability, but such an approach requires a protein structure, which is, however, not always available for most proteins (Bakheet and Doig, 2009). If the protein structure is unavailable druggability can be "predicted" by comparing the amino acid sequence of interest with that of known drug targets, in the StARLITe/ChEMBL database, using Basic Local Alignment Search Tool (BLAST). According to Crowther *et al.* (2010), a protein can be considered druggable if:

1. It is $\geqslant 80\%$ of the length of the corresponding druggable target;
2. It has an amino acid sequence that aligns with $\geqslant 80\%$ of the druggable target or,
3. The BLAST expectation value of the alignment is less than $10e^{-10}$ (Crowther *et al.*, 2010).

1.2.1.3 Is the protein essential for parasite survival?

Putative targets must be essential for the survival of the pathogen (Perumal *et al.*, 2009).

Load points

"Load points" can be defined as "*hotspots in a metabolic network based on the ratio of number of k-shortest paths passing through the metabolite/enzyme (in/out) and the number of the nearest neighbour links (in/out) attached to it, compared to the average load value in the network*". These load point values give a global view of the metabolic network and help in the analysis of the metabolic pathway reactions. Pathways that are highly connected in the metabolism of the cell tend to have high load values. Also the lethality of a metabolite/enzyme depends on the number of connections it has in the whole metabolic network. Enzymes with more connections are found to be more essential than the proteins that interact with only a few neighbours (Perumal *et al.*, 2009; Rahman and Schomburg, 2006). Recently it has been found that proteins with a high load are mostly housekeeping proteins and are thus also conserved across all organisms making them not ideal drug targets (Chanumolu *et al.*, 2012). For a given metabolic network, the load "L" on metabolite "m" can be calculated as in Equation

1.1; where $-\infty < \mathrm{Lm(in/out)} < \infty$, p is the number of shortest paths (in/out) passing through a metabolite m; k is the number of nearest neighbour links (in/out) for m in the network; P is the total number of shortest paths and K is the sum of links in the metabolic network of M metabolites (where M is the number of metabolites in the network) (Rahman and Schomburg, 2006):

$$L_{m(in/out)} = ln[(P_{m(in/out)}/k_{m(in/out)})/(\sum_{i=1}^{M} Pi(in/out))/\sum_{i=1}^{M} Ki_{(in/out)}] \tag{1.1}$$

Chokepoints

A "chokepoint" can be defined as "*a reaction that either uniquely consumes a specific substrate or uniquely produces a specific product in the metabolic pathway*" (Hasan *et al.*, 2006). Chokepoint enzymes are crucial points in the metabolic pathway and inactivation of these enzymes may lead to the disruption of the metabolic network of a pathogen (Perumal *et al.*, 2009), because their function cannot be compensated for by another enzyme, making them good metabolic drug targets (Hasan *et al.*, 2006). An extended graph theory model ranks the chokepoints according to the k-shortest path passing through it and the load (in/out) on it. This ranking has an advantage as this measure may help determine the biochemical essentiality of a metabolite/enzyme as chokepoints may not be essential if they create unique intermediates to an essential product and exhibit alternate pathway reactions (Perumal *et al.*, 2009; Rahman and Schomburg, 2006).

End metabolite production within a pathway normally depends upon the sequential completion of all the reactions in a pathway and the supporting reactions from other pathways. Thus, by inhibiting any enzyme in the upstream regulating reactions would lead to a discontinuation of its production which can lead to a disturbances in various dependent metabolic pathways which use the given end metabolite as a substrate. By inhibiting a unique chokepoint enzyme it could lead to either accumulation of the unique substrate, which can be potentially toxic to the cell, or starvation of the unique product, which can be potentially crippling for essential cell functions, as no other shunt path would be available to complete the reaction (Chanumolu *et al.*, 2012).

1.2.1.4 Does the protein have a human homolog?

Enzymes that do not have human homologs may be attractive drug targets, thus a comparative study between the human metabolic network and the parasite metabolic network are essential to identify possible interference of the drugs with the human metabolism as this might lead to possible side effects (Florez *et al.*, 2010; Hasan *et al.*, 2006; Perumal *et al.*, 2009; Rahman and Schomburg, 2006). An ideal drug against parasites or other disease-causing agents would have minimal interaction with the host while also having high binding specificity for the pathogen.

To increase the chances of this favourable binding pattern, targets can be penalized for having high sequence similarity to its host (Hasan *et al.*, 2006) or rewarded if there are no orthologs between the host and the pathogen. But it has to be kept in mind, though, that presently a large number of genes in most organisms have unidentified functions which could affect erroneous predictions (Rahman and Schomburg, 2006).

1.2.1.5 Is the protein active in the diseased state?

For a parasite protein to be available for drug-target interactions the parasite protein target needs to be present in the host (Hasan *et al.*, 2006). If a parasite protein is not expressed/active in the host it would not be an ideal drug target as the parasite protein on which the drug should bind would be absent from the host, while if a parasite protein is expressed/active in the host it would be available for binding to a drug, and it would thus be possible to inactivate the parasite protein that is needed for the parasite's survival in the host.

If a parasite protein target is expressed/active in a latent *in vivo* microarray model it increases the confidence that it is expressed in latent *in vivo* infection (Hasan *et al.*, 2006), increasing the attraction of using it as a drug target.

1.2.2 Scoring

Target prioritization is probably the most useful as a prelude, rather than a replacement, of laborious experimental follow-up work. Experimental characterization of promising targets often requires chemical inhibitors of target activity; therefore lists of target-specific inhibitors would be of great value to the research community (Crowther *et al.*, 2010).

It is vital to have as much evidence as possible to support a target choice before investing more resources in the target (Bakheet and Doig, 2009). By scoring the targets the most promising targets might be identified.

1.2.2.1 Boolean intersection

In multi-criteria searches it is possible to take a Boolean intersection of the criteria so that only those proteins with all the desired traits are selected (Crowther *et al.*, 2010; Khoshkholgh-Sima *et al.*, 2011). This is a high risk as a protein may lack one or more preferred property and still be an effective drug target (Crowther *et al.*, 2010; Khoshkholgh-Sima *et al.*, 2011). These queries will eliminate targets that should be considered but which fail to meet one or a few of the specified criteria. The incompleteness of available experimental data means that in many cases intersection queries may be too stringent for prioritizing drug targets (Aguero *et al.*, 2008).

Kumar *et al.* (2007) used this method of scoring for predicting drug targets in *Brugia malayi*, and Caffrey *et al.* (2009), for predicting drug targets in *Schistosoma mansoni (Crowther* et al.*, 2010).*

1.2.2.2 Weighted scores

A criterion is assigned a subjective weight and targets earn points for each criterion they meet (Crowther *et al.*, 2010), and how well they are met for some criterion. Important criteria are given large weights and might include: growth essentiality, druggable protein domain, metabolic chokepoint, availability of structural clues, and absence of close homolog and gene expression in disease model (Hasan *et al.*, 2006). Less important criteria are given small weights while undesired criterions are given negative weights (Crowther *et al.*, 2010).

These queries would return a ranked list of all potential targets, ordered by a cumulative score. Targets can be re-ranked by changing the weights and/or adding additional criteria (Crowther *et al.*, 2010). This list is imperfect, but might suggest rational choices due to the plausible and informative rankings of the target across the criteria (Crowther *et al.*, 2010).

In the Tropical Disease Research database (TDR) (for more information regarding TDR see Section 1.4.5) a user chooses his/her own scores to generate a ranking according to the user's preferences, individual queries are each assigned a numerical weight by the user and then combined, with the end list ranked according to the additive weighting. The user assigns higher numbers to features considered to be particularly important and lower numbers to features that are desirable but not indispensable (Aguero *et al.*, 2008).

The weighted approach will avoid premature elimination of potential drug targets.

1.2.3 Results

1.2.3.1 Old vs. New

Having well-known targets, or targets of known drugs, at the top of the ranked target list offers assurance that the search strategy are reasonable and can act as controls, but a method that only identifies well-established targets would not serve the important purpose of suggesting new targets. Thus the presence of new, or even hypothetical, targets near the top is also sought (Crowther *et al.*, 2010).

1.2.3.2 False results

False negatives

False negatives will result when a known or well-established target ranks low on the list of possible targets. This might be due to several reasons that might include: difficulty of assaying the protein in isolation, drugs might have been identified through phenotypic screens and the target do not meet many of the required criteria in a target based approach, or it is assumed that drugs will cause a loss-of-function where it might cause a gain-of-function (Crowther *et al.*, 2010).

False negative results can be discouraging, but adding additional datasets can potentially eliminate this (Crowther *et al.*, 2010).

False positives

False positives are when proteins rank highly but could not be validated as a drug target, despite considerable efforts (Crowther *et al.*, 2010). Limitations in scoring methods can result in high false positive prediction rates. False positives can potentially be eliminated by changing the scoring strategy and also by adding additional datasets.

1.2.4 Validation

Genome projects are generating a large number of potential new targets for drug discovery. But one challenge is target validation, providing the usefulness of a specific target in an animal model (Tao *et al.*, 2000). Selecting the most useful target is limiting the need to define those targets essential to the disease process in an animal model, or cell culture, and by the lack of biochemical characterization of many of them (Tao *et al.*, 2000). Technologies currently available to validate targets include:

- Gene knockouts (Tao *et al.*, 2000; Torrie *et al.*, 2009);
- RNA interference (Torrie *et al.*, 2009);
- Signature-tagged mutagenesis (Tao *et al.*, 2000);
- Genomic analysis and mapping by *in vitro* transposition (Tao *et al.*, 2000);
- Antisense approaches (Tao *et al.*, 2000) or,
- Temperature-sensitive mutants (Tao *et al.*, 2000).

These are not designed to test whether the wild-type protein target is essential for pathogen growth in the disease state. In particular, it is desirable to demonstrate the reversal of the disease state by modulating the function of the target *in vivo* (Tao *et al.*, 2000).

1.3 Approaches

Strategies usually aim to identify parasite-specific enzymes, complete biochemical pathways or structural differences between human and parasite homologues. The latter approach is hampered by the limited availability of 3D structures (Alves-Ferreira *et al.*, 2009). Several strategies exist for the pursuit of drugs to treat diseases. Major approaches can normally be classified as (a) **label extension**, extending the indications of existing drugs for other condition to new diseases; (b) **piggy-back discovery**, the discovery of new drugs is focused on one or a few classes of well-studied and validated targets; and (c) ***de novo* drug discovery** (Crowther *et al.*, 2010).

1.3.1 Repurposing

Drug repurposing is when new therapeutic indications; like treating diseases other than their original indications; are found for old drugs (Liu *et al.*, 2010; Yang *et al.*, 2013), it can also be called label extension (Crowther *et al.*, 2010).

By screening drugs that are already approved by the FDA it can increase the speed of the clinical trial and drug approval process as it has been found to be already safe for human consumption (Yang *et al.*, 2013), it only needs to be demonstrated that it can be used effectively for a different indication (Karczewski *et al.*, 2012), thus, reducing cost and time spent in clinical trials (Yang *et al.*, 2013). Any method that can be used to characterize "off-target" effects can be used in drug repurposing, by finding effects that are favourable for a disease being studied (Karczewski *et al.*, 2012).

Either receptor-based or ligand-based *in silico* methods have been applied to drug repurposing projects. Phatak and Zhang (2013) states that there was a case of 23 predicted and validated novel drug-target associations using two dimensional (2D) chemical similarity approaches (Phatak and Zhang, 2013).

Some examples might include:

- A Parkinson's disease drug for Tuberculosis:
  - It has been shown that a drug target for Parkinson's disease, catchol-O-methyltransferase (COMT) and a bacterial protein in *Mycobacterium tuberculosis* (*M. tuberculosis*) (the enoyl-acyl carrier protein reductase, InhA) are very similar. This has narrowed down an investigation of potential drug targets for *M. tuberculosis* infections. From this, Entacapone, a drug already approved to treat Parkinson's by inhibiting COMT, was predicted to bind to InhA (Karczewski *et al.*, 2012).
- A chronic myelogenous leukemia drug for prostate and breast cancer:
  - Activated CDC42 kinase 1 (ACK1), is a novel cancer target that gets significantly overexpressed in breast and prostate cancers during its growth. It also regulates several downstream proteins that have been implicated in cell survival roles. By comparing this protein to various drugs a short list of 10 drugs was identified. Of these Dasatinib was selected as it had reasonable interactions with ACK1, based on docking studies. It also performed well in experimental studies (Phatak and Zhang, 2013).

1.3.2 Chemogenomics

Chemogenomics can be applied to a wide range of approaches that uses chemical compounds to probe biological systems. All of the approaches have some relevance to drug discovery, but they can differ according to the extent to which they employ stochastic versus directed approaches (Mannhold *et al.*, 2004).

The stochastic chemogenomic approach probes the global response of a biological system on exposure to chemical compounds while the focused chemogenomic approaches uses chemicals as detailed probes of biochemical pathways that can play a key role in target identification and validation (Mannhold *et al.*, 2004).

An integrated chemogenomics platform uses affinity-based screening, directed combinatorial chemistry and SBDD to develop drug-like tool compounds that can validate a target-based therapeutic hypothesis *in vivo* (Mannhold *et al.*, 2004). The objective of chemogenomics is to find and optimize chemical compounds that can be used to directly test the therapeutic relevance of new targets revealed through genome sequencing (Mannhold *et al.*, 2004).

Chemogenomics has emerged in target prediction via data mining in target-annotated databases. The success of this technique depends on the availability of bioactivity data for the targets and their associated ligands. For new ligands this data is normally only approximate or unavailable in lack of corresponding target information (Liu *et al.*, 2010).

1.3.3 Reverse Docking

Another *in silico* target prediction method is reverse docking. Reverse docking uses a given small molecule and then probes the potential ligand binding sites (Liu *et al.*, 2010) of all the 3D structures in a given protein database (Li *et al.*, 2006). Protein "hits" can then serve as potential candidates for experimental validation (Li *et al.*, 2006).

The small molecule might be a biologically active compound detected in a cell- or animal-based bioassay screen, a natural product or an existing drug whose molecular target(s) is (are) unidentified (Li *et al.*, 2006; Liu *et al.*, 2010).

1.3.4 Evolutionary tracing

A number of bioinformatics approaches may be used to identify essential amino acids in potential drug targets. One approach, named evolutionary tracing (ET), makes use of protein multiple sequence alignments (MSA) (Durand *et al.*, 2008). ET is based on the hypothesis that architecture-defining residues are mostly constant, and traces these residues through a phylogenetic tree to guide investigators to structurally relevant sites. But, these protein homology-based methods are, however, limited since some functional regions involve contact areas that may only be apparent from 3D protein structures and are not obvious from primary sequence alignment. Functional regions can also be organism specific, particularly if the sequence homologies are low, and may not be clear from protein MSA (Durand *et al.*, 2008).

Homology methods do not include the evolutionary information available from nucleic acid sequences. It has been acknowledged that alternative approaches, such as those that make use of evolutionary analyses, should be applied at various points in the drug development pipeline. This is particularly appropriate for pathogen drug design since whole genome data from several parasites are available (Durand *et al.*, 2008).

1.3.5 Evolution patterning

"Evolutionary patterning" (EP) aims to identify and assess the suitability of a potential drug target sites from the point of view of limiting the evolution of mutations conferring drug resistance (Durand *et al.*, 2008). EP makes use of the pattern of evolutionary change at individual codons across coding sequences to limit drug resistance by identifying the most constrained sites. EP can be combined with structural information and, like ET, is generic in nature. Structural modelling was used to refine the selection of potential target sites by assessing their functional and structural significance (Durand *et al.*, 2008).

Nucleotide sequences contain information about the rate of evolution, which can be measured to determine the intensity of the selective force acting at a particular site in a protein. If a particular residue is critical to the structure or function of a protein, natural selection will remove any changes that occur at that site (purifying selection) at a rate that reflects its relative importance. These changes produce a pattern of evolution. Since evolutionary pressure act to maintain the most critical residues, it follow that mutations that confer drug resistance are unlikely to evolve at these sites (Durand *et al.*, 2008).

If a potential target is under extreme purifying selection these residues could be selected as drug target sites. It should be noted that this type of analysis represents an initial step, since druggability of each site requires evaluation by medicinal chemists (Durand *et al.*, 2008).

1.4 Databases and web resources

A target database may provide not only abundant information about the potential target protein; such as 3D structures, binding (active) sites, biological (pharmacological) functions and related diseases; but also appropriate computational tools to mine the information about targets (Gao *et al.*, 2008). The following databases and web resources were chosen as they looked promising to predict ligand or drug binding sites by using a 2D protein structure, and does not necessarily rely on a 3D structure as well to generate a list of possible targets according to the user's specifications.

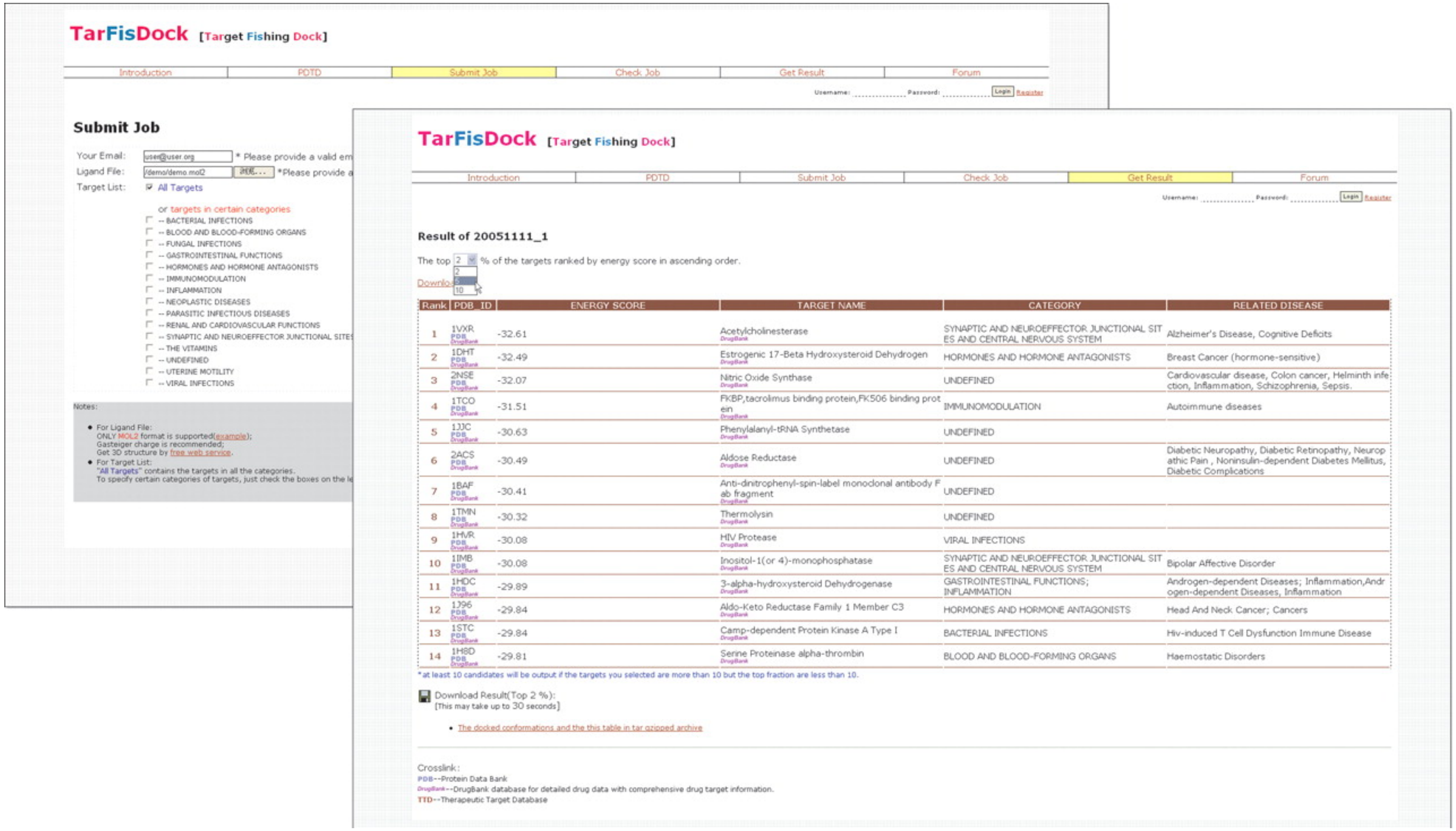
1.4.1 Target Fishing Dock

Target Fishing Dock (TarFisDock) seeks potential binding proteins for a given ligand. It makes use of a ligand-protein reverse docking strategy to search out all possible proteins for a small molecule from the potential drug target database. TarFisDock might serve as a valuable tool for identifying targets for a novel synthetic compound or for a newly isolated natural product, for a compound with known biological activity, or for an existing drug whose mechanism is unknown (Li *et al.*, 2006).

TarFisDock consist of two parts, a front-end web interface and a back-end tool for reverse docking. TarFisDock was developed on the basis of the widely used docking program, DOCK. The reverse docking procedure is as follows:

- TarFisDock generates a protein target list according to the user's preference or selects all the protein entries in the PDTD if the user intends to find a new target for an active compound;
- TarFisDock docks a given small molecule into the possible binding sites of proteins in the target list, and the interaction energies between the small molecule and the proteins are calculated and recorded;
- TarFisDock analyses the reverse docking result (Li *et al.*, 2006).

The input file of TarFisDock consists of only the test small molecule in the mol2 format. The 2D structure of a small molecule can be either sketched or taken from chemical databases. The user can convert the small molecule from its 2D structures to the 3D structure (Figure 1.3). The final 3D structure is saved in a mol2 file (Li *et al.*, 2006).



Submit Job

Your Email: user@user.org * Please provide a valid em

Ligand File: /demo/demo.mol2 *Please provide a

Target List: ☑ All Targets

or targets in certain categories

- ☐ -- BACTERIAL INFECTIONS
- ☐ -- BLOOD AND BLOOD-FORMING ORGANS
- ☐ -- FUNGAL INFECTIONS
- ☐ -- GASTROINTESTINAL FUNCTIONS
- ☐ -- HORMONES AND HORMONE ANTAGONISTS
- ☐ -- IMMUNOMODULATION
- ☐ -- INFLAMMATION
- ☐ -- NEOPLASTIC DISEASES
- ☐ -- PARASITIC INFECTIOUS DISEASES
- ☐ -- RENAL AND CARDIOVASCULAR FUNCTIONS
- ☐ -- SYNAPTIC AND NEUROEFFECTOR JUNCTIONAL SITES
- ☐ -- THE VITAMINS
- ☐ -- UNDEFINED
- ☐ -- UTERINE MOTILITY
- ☐ -- VIRAL INFECTIONS

Notes:

- For Ligand File:
  ONLY MOL2 format is supported(example);
  Gasteiger charge is recommended;
  Get 3D structure by free web service.
- For Target List:
  "All Targets" contains the targets in all the categories.
  To specify certain categories of targets, just check the boxes on the le

Result of 20051111_1

The top 2 % of the targets ranked by energy score in ascending order.

Downlo

Rank	PDB_ID	ENERGY SCORE	TARGET NAME	CATEGORY	RELATED DISEASE
1	1VXR PDB DrugBank	-32.61	Acetylcholinesterase DrugBank	SYNAPTIC AND NEUROEFFECTOR JUNCTIONAL SITES AND CENTRAL NERVOUS SYSTEM	Alzheimer's Disease, Cognitive Deficits
2	1DHT PDB DrugBank	-32.49	Estrogenic 17-Beta Hydroxysteroid Dehydrogen DrugBank	HORMONES AND HORMONE ANTAGONISTS	Breast Cancer (hormone-sensitive)
3	2NSE PDB DrugBank	-32.07	Nitric Oxide Synthase DrugBank	UNDEFINED	Cardiovascular disease, Colon cancer, Helminth infection, Inflammation, Schizophrenia, Sepsis.
4	1TCO PDB DrugBank	-31.51	FKBP,tacrolimus binding protein,FK506 binding protein DrugBank	IMMUNOMODULATION	Autoimmune diseases
5	1JJC PDB DrugBank	-30.63	Phenylalanyl-tRNA Synthetase DrugBank	UNDEFINED	
6	2ACS PDB DrugBank	-30.49	Aldose Reductase DrugBank	UNDEFINED	Diabetic Neuropathy, Diabetic Retinopathy, Neuropathic Pain , Noninsulin-dependent Diabetes Mellitus, Diabetic Complications
7	1BAF PDB DrugBank	-30.41	Anti-dinitrophenyl-spin-label monoclonal antibody Fab fragment DrugBank	UNDEFINED	
8	1TMN PDB DrugBank	-30.32	Thermolysin DrugBank	UNDEFINED	
9	1HVR PDB DrugBank	-30.08	HIV Protease DrugBank	VIRAL INFECTIONS	
10	1IMB PDB DrugBank	-30.08	Inositol-1(or 4)-monophosphatase DrugBank	SYNAPTIC AND NEUROEFFECTOR JUNCTIONAL SITES AND CENTRAL NERVOUS SYSTEM	Bipolar Affective Disorder
11	1HDC PDB DrugBank	-29.89	3-alpha-hydroxysteroid Dehydrogenase DrugBank	GASTROINTESTINAL FUNCTIONS; INFLAMMATION	Androgen-dependent Diseases; Inflammation,Androgen-dependent Diseases, Inflammation
12	1J96 PDB DrugBank	-29.84	Aldo-Keto Reductase Family 1 Member C3 DrugBank	HORMONES AND HORMONE ANTAGONISTS	Head And Neck Cancer; Cancers
13	1STC PDB DrugBank	-29.84	Camp-dependent Protein Kinase A Type I DrugBank	BACTERIAL INFECTIONS	Hiv-induced T Cell Dysfunction Immune Disease
14	1H8D PDB DrugBank	-29.81	Serine Proteinase alpha-thrombin DrugBank	BLOOD AND BLOOD-FORMING ORGANS	Haemostatic Disorders

*at least 10 candidates will be output if the targets you selected are more than 10 but the top fraction are less than 10.

Download Result(Top 2 %):
[This may take up to 30 seconds]

- The docked conformations and the this table in tar gzipped archive

Crosslink:
PDB--Protein Data Bank
DrugBank--DrugBank database for detailed drug data with comprehensive drug target information.
TTD--Therapeutic Target Database

Figure 1.3: An example of the input and output of TarFisDock.

The input file of TarFisDock is the test small molecule in a mol2 format. The output is delivered in ascending order of energy score. The archived file contains a list of all the scores, as well as the binding models of the small molecule tested within the binding sites of the candidate targets.

From: Li *et al.* (2006).

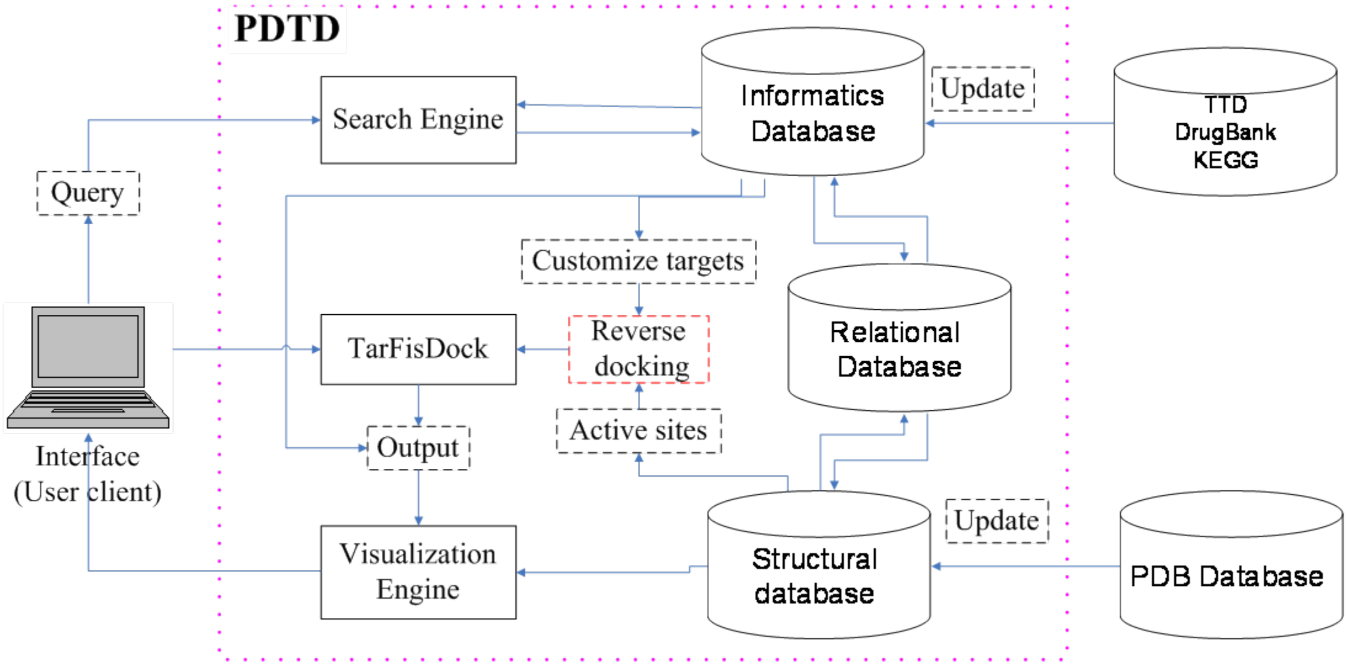


Figure 1.4: PDTD system architecture.

PDTD contains two sub-databases types; the structural sub-database and the informatics sub-database. All data are associated with a relational database implemented using MySQL and can be queried through the web interface. Via three computational engines: search engine, visualization engine and TarFisDock, users can implement an interactive query and computation with the PDTD. The structural sub-database stores each protein in both PDB format and mol2 format with Amber charges. Sequence and active site information were also included in the structural sub-database. The informatics database stores the data of target categories, related disease information, biological functions and associated regulating pathways.

From: Gao *et al.* (2008).

The output is delivered in ascending order of energy score (Figure 1.3). The archive file contains a list of the scores, together with binding models of the small molecule tested within the binding sites of the candidate targets (Li *et al.*, 2006).

TarFisDock has limitations, these include:

- Protein entries are not enough for covering all the protein information of disease-related genomes;
- TarFisDock has not considered the flexibility of proteins during docking simulation. These two limitations will produce false negatives. And,
- The scoring function for reverse docking is not accurate enough. This will produce false positives (Li *et al.*, 2006).

1.4.2 Potential Drug Target Database

Potential Drug Target Database (PDTD) is a valuable platform for target identification and can be integrated with TarFisDock (Gao *et al.*, 2008). PDTD is a comprehensive, web-accessible database of drug targets, which focuses on those drug targets with known 3D-structure. Drug targets of PDTD were categorized by two criteria: therapeutic areas and biochemical criteria. Each target was carefully annotated by browsing several databases, such as DrugBank and UniProt (Gao *et al.*, 2008).

PDTD has dual functions querying drug target information and identifying the potential binding protein of an active compound or an existing drug by using reverse docking approach. Accordingly, PDTD contains two sub-databases types, one is the structural sub-database and the other is the informatics sub-database (Figure 1.4) (Gao *et al.*, 2008). Through three computational engines, search engine, visualization engine and TarFisDock, users can implement interactive query and computation with the PDTD. The structural sub-database stores each protein in both Protein Databank (PDB) format and mol2 format with Amber charges; sequence and active site information were also included in the structural sub-database. The informatics database stores the data of target categories, related disease information, biological functions and associated regulating pathways (Gao *et al.*, 2008).

Users can also use reverse docking to search PDTD for finding the possible binding proteins(s) of a small molecule (Gao *et al.*, 2008). PDTD is supported with a friendly designed web interface so that users can easily query the target information, and retrieve, visualized or download the distributions of the drug target files, as they desire. Figure 1.5 shows screenshots to illustrate some of the information describing the drug target. These views include the home view of PDTD that introduces this database; the Browse view, which is subdivided into Therapeutic Areas, Biochemical Type, PDB ID and Target Name. The database can be browsed through any of these means; Search view where a Target Name, PDB ID or Disease Name can be used as search text to search the database. The last screen shot displays detail of a certain protein that can be view by choosing a protein after a search, or after browsing the database (Gao *et al.*, 2008).

Every target has its own result page containing comprehensive information including PDB ID, target name, target category, related disease, its structure, and active site. The PDTD was carefully annotated according to information found in the PDB, UniProt, KEGG, and Enzyme Structure Database (Gao *et al.*, 2008).

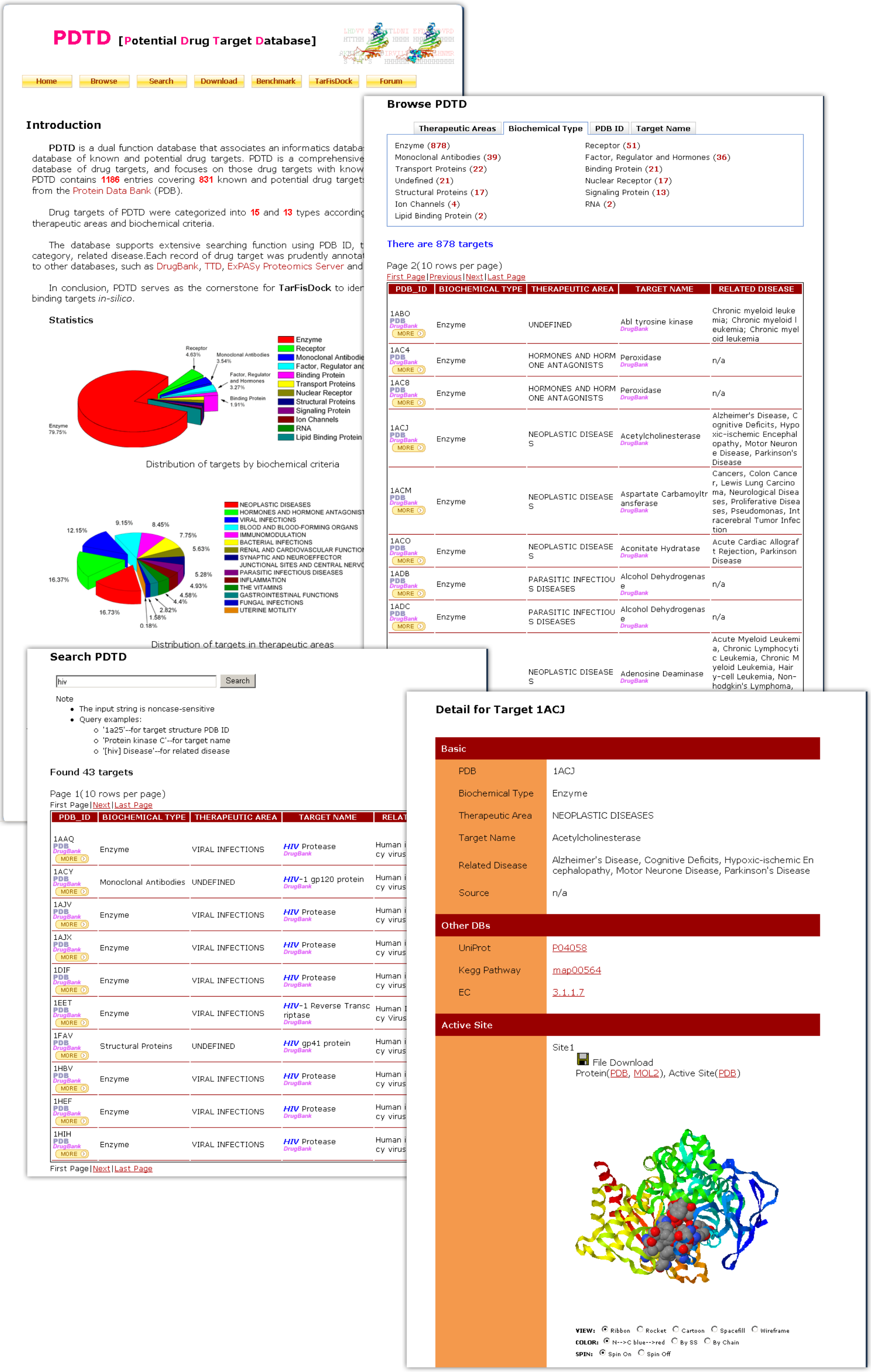
1.4.3 PharmMapper

The abundant potential target entries represented by pharmacophore models in PharmTargetDB and the efficient pharmacophore-mapping algorithm allows fast and reliable identification of the pharmacophore target candidates for small molecules such as drugs, lead compounds and natural products (Liu *et al.*, 2010). PharmMapper, an open-access tool, can provide useful clues for further bioassay in drug-target interaction research (Liu *et al.*, 2010).

Similar to TarFisDock, PharmMapper can also be used in mapping the regulation genomic network for an existing drug of a drug candidate, as well as in profiling the potential secondary or side effects for a drug molecule in a different viewpoint from the regular chemogenetic method (Liu *et al.*, 2010). PharmMapper bears high throughput ability and can identify the potential target candidates from the database with a runtime of a few hours. This target information can be significant for functional genomic study within the chemical biology paradigm (Liu *et al.*, 2010).

PharmMapper consists of two parts: a front-end web interface and a back-end tool for reverse pharmacophore mapping (Liu *et al.*, 2010). It only needs a file with a single drug-like molecule or natural product as input (Liu *et al.*, 2010). Figure 1.6 shows an example of the output of PharmMapper. The output of a PharmMapper run is demonstrated in the form of a ranked list of hit target pharmacophore models that are sorted by fit score in descending order (Figure 1.6A). Result lists can be re-ranked by normalized fit score or number of pharmacophore features in descending order via clicking the arrow icons in the corresponding columns.

Figure 1.6B shows that when the '+' at the start of each line in the result table is clicked the details



PDTD [Potential Drug Target Database]

Introduction

PDTD is a dual function database that associates an informatics databa database of known and potential drug targets. PDTD is a comprehensive database of drug targets, and focuses on those drug targets with know PDTD contains 1186 entries covering 831 known and potential drug target from the Protein Data Bank (PDB).

Drug targets of PDTD were categorized into 15 and 13 types according therapeutic areas and biochemical criteria.

The database supports extensive searching function using PDB ID, t category, related disease.Each record of drug target was prudently annotat to other databases, such as DrugBank, TTD, ExPASy Proteomics Server and

In conclusion, PDTD serves as the cornerstone for TarFisDock to ider binding targets *in-silico*.

Statistics

- 🟥 Enzyme
- 🟩 Receptor
- 🟦 Monoclonal Antibodie
- Factor, Regulator anc
- Binding Protein
- 🟨 Transport Proteins
- Nuclear Receptor
- Structural Proteins
- 🟪 Signaling Protein
- Ion Channels
- 🟩 RNA
- Lipid Binding Protein

Distribution of targets by biochemical criteria

- 🟥 NEOPLASTIC DISEASES
- 🟩 HORMONES AND HORMONE ANTAGONIST
- 🟦 VIRAL INFECTIONS
- BLOOD AND BLOOD-FORMING ORGANS
- IMMUNOMODULATION
- 🟨 BACTERIAL INFECTIONS
- RENAL AND CARDIOVASCULAR FUNCTIOI
- SYNAPTIC AND NEUROEFFECTOR JUNCTIONAL SITES AND CENTRAL NERVC
- 🟪 PARASITIC INFECTIOUS DISEASES
- INFLAMMATION
- 🟩 THE VITAMINS
- GASTROINTESTINAL FUNCTIONS
- 🟦 FUNGAL INFECTIONS
- 🟧 UTERINE MOTILITY

Distribution of targets in therapeutic areas

Browse PDTD

Therapeutic Areas | Biochemical Type | PDB ID | Target Name

Enzyme (878)	Receptor (51)
Monoclonal Antibodies (39)	Factor, Regulator and Hormones (36)
Transport Proteins (22)	Binding Protein (21)
Undefined (21)	Nuclear Receptor (17)
Structural Proteins (17)	Signaling Protein (13)
Ion Channels (4)	RNA (2)
Lipid Binding Protein (2)	

There are 878 targets

Page 2(10 rows per page)
First Page|Previous|Next|Last Page

PDB_ID	BIOCHEMICAL TYPE	THERAPEUTIC AREA	TARGET NAME	RELATED DISEASE
1ABO PDB DrugBank MORE	Enzyme	UNDEFINED	Abl tyrosine kinase DrugBank	Chronic myeloid leukemia; Chronic myeloid leukemia; Chronic myeloid leukemia
1AC4 PDB DrugBank MORE	Enzyme	HORMONES AND HORMONE ANTAGONISTS	Peroxidase DrugBank	n/a
1AC8 PDB DrugBank MORE	Enzyme	HORMONES AND HORMONE ANTAGONISTS	Peroxidase DrugBank	n/a
1ACJ PDB DrugBank MORE	Enzyme	NEOPLASTIC DISEASES	Acetylcholinesterase DrugBank	Alzheimer's Disease, Cognitive Deficits, Hypoxic-ischemic Encephalopathy, Motor Neurone Disease, Parkinson's Disease
1ACM PDB DrugBank MORE	Enzyme	NEOPLASTIC DISEASES	Aspartate Carbamoyltransferase DrugBank	Cancers, Colon Cancer, Lewis Lung Carcinoma, Neurological Diseases, Proliferative Diseases, Pseudomonas, Intracerebral Tumor Infection
1ACO PDB DrugBank MORE	Enzyme	NEOPLASTIC DISEASES	Aconitate Hydratase DrugBank	Acute Cardiac Allograft Rejection, Parkinson Disease
1ADB PDB DrugBank MORE	Enzyme	PARASITIC INFECTIOUS DISEASES	Alcohol Dehydrogenase DrugBank	n/a
1ADC PDB DrugBank MORE	Enzyme	PARASITIC INFECTIOUS DISEASES	Alcohol Dehydrogenase DrugBank	n/a
		NEOPLASTIC DISEASES	Adenosine Deaminase DrugBank	Acute Myeloid Leukemia, Chronic Lymphocytic Leukemia, Chronic Myeloid Leukemia, Hairy-cell Leukemia, Non-hodgkin's Lymphoma,

Search PDTD

hiv Search

Note

- The input string is noncase-sensitive
- Query examples:
  - '1a25'--for target structure PDB ID
  - 'Protein kinase C'--for target name
  - '[hiv] Disease'--for related disease

Found 43 targets

Page 1(10 rows per page)
First Page|Next|Last Page

PDB_ID	BIOCHEMICAL TYPE	THERAPEUTIC AREA	TARGET NAME	RELAT
1AAQ PDB DrugBank MORE	Enzyme	VIRAL INFECTIONS	*HIV* Protease DrugBank	Human i cy virus
1ACY PDB DrugBank MORE	Monoclonal Antibodies	UNDEFINED	*HIV*-1 gp120 protein DrugBank	Human i cy virus
1AJV PDB DrugBank MORE	Enzyme	VIRAL INFECTIONS	*HIV* Protease DrugBank	Human i cy virus
1AJX PDB DrugBank MORE	Enzyme	VIRAL INFECTIONS	*HIV* Protease DrugBank	Human i cy virus
1DIF PDB DrugBank MORE	Enzyme	VIRAL INFECTIONS	*HIV* Protease DrugBank	Human i cy virus
1EET PDB DrugBank MORE	Enzyme	VIRAL INFECTIONS	*HIV*-1 Reverse Transcriptase DrugBank	Human I cy Virus
1FAV PDB DrugBank MORE	Structural Proteins	UNDEFINED	*HIV* gp41 protein DrugBank	Human i cy virus
1HBV PDB DrugBank MORE	Enzyme	VIRAL INFECTIONS	*HIV* Protease DrugBank	Human i cy virus
1HEF PDB DrugBank MORE	Enzyme	VIRAL INFECTIONS	*HIV* Protease DrugBank	Human i cy virus
1HIH PDB DrugBank MORE	Enzyme	VIRAL INFECTIONS	*HIV* Protease DrugBank	Human i cy virus

First Page|Next|Last Page

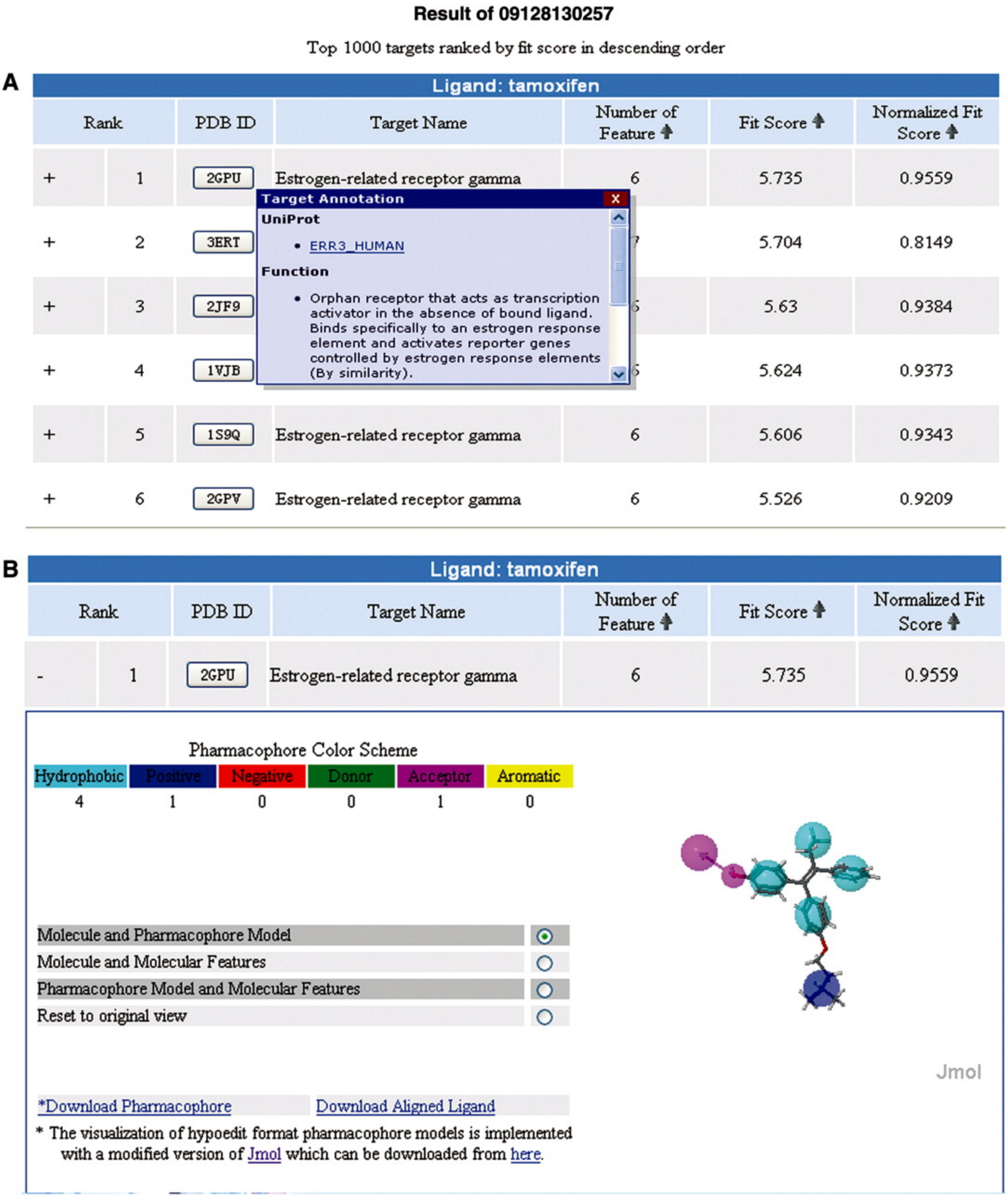
Detail for Target 1ACJ

Basic	
PDB	1ACJ
Biochemical Type	Enzyme
Therapeutic Area	NEOPLASTIC DISEASES
Target Name	Acetylcholinesterase
Related Disease	Alzheimer's Disease, Cognitive Deficits, Hypoxic-ischemic Encephalopathy, Motor Neurone Disease, Parkinson's Disease
Source	n/a
Other DBs	
UniProt	P04058
Kegg Pathway	map00564
EC	3.1.1.7
Active Site	
	Site1 File Download Protein(PDB, MOL2), Active Site(PDB)

Figure 1.5: Screen shots of the PDTD.

A screen shot of the PDTD showing several possible view of information describing the drug target. Not all the fields are shown. PDTD has been designed to provide fast and easy access to target information.

From: Gao *et al.* (2008).



Result of 09128130257

Top 1000 targets ranked by fit score in descending order

Ligand: tamoxifen

Rank		PDB ID	Target Name	Number of Feature	Fit Score	Normalized Fit Score
+	1	2GPU	Estrogen-related receptor gamma	6	5.735	0.9559
+	2	3ERT		7	5.704	0.8149
+	3	2JF9		6	5.63	0.9384
+	4	1VJB		6	5.624	0.9373
+	5	1S9Q	Estrogen-related receptor gamma	6	5.606	0.9343
+	6	2GPV	Estrogen-related receptor gamma	6	5.526	0.9209

Target Annotation

UniProt

- ERR3_HUMAN

Function

- Orphan receptor that acts as transcription activator in the absence of bound ligand. Binds specifically to an estrogen response element and activates reporter genes controlled by estrogen response elements (By similarity).

Ligand: tamoxifen

Rank		PDB ID	Target Name	Number of Feature	Fit Score	Normalized Fit Score
-	1	2GPU	Estrogen-related receptor gamma	6	5.735	0.9559

Pharmacophore Color Scheme

Hydrophobic	Positive	Negative	Donor	Acceptor	Aromatic
4	1	0	0	1	0

Molecule and Pharmacophore Model ●
Molecule and Molecular Features ○
Pharmacophore Model and Molecular Features ○
Reset to original view ○

*Download Pharmacophore Download Aligned Ligand

* The visualization of hypoedit format pharmacophore models is implemented with a modified version of Jmol which can be downloaded from here.

Figure 1.6: Output of PharmMapper.

(A) The ranked list of hit target pharmacophore models, which are sorted by fit score in descending order. (B) The pull-down window that illustrates the details of each pharmacophore model candidate and the molecule pharmacophore alignment.

From: Liu *et al.* (2010).

of each pharmacophore model candidate is illustrated, including the numbers of each pharmacophore feature, a 3D interactive visualization of molecule-pharmacophore alignment poses, and the download links of the aligned pose of molecule as well as the corresponding pharmacophore model. The radio buttons in the pull down window allow the users to show/hide either the pharmacophore model; query molecular conformation or the features form the query molecule in display (Liu *et al.*, 2010).

1.4.4 DrugBank

DrugBank is a dual-purpose bioinformatics-chemoinformatics database with a strong focus on quantitative, analytic or molecular-scale information about drugs and drug targets. It combines the data-rich molecular biology content normally found in curated sequence databases with the equally rich data found in medicinal chemistry textbooks and chemical reference handbooks (Wishart *et al.*, 2006). The aim of DrugBank is to allow educators and researchers from diverse backgrounds and disciplines to conduct the type of *in silico* learning and discovery that is now routine in the world of genomics and proteomics (Wishart *et al.*, 2006).

The assembly of DrugBank was difficult and time consuming as most of the data were "paperbound". DrugBank currently contains $> 4,199$ drug entries, corresponding to $> 12,000$ different trade names and synonyms. These drug entries were chosen on the following:

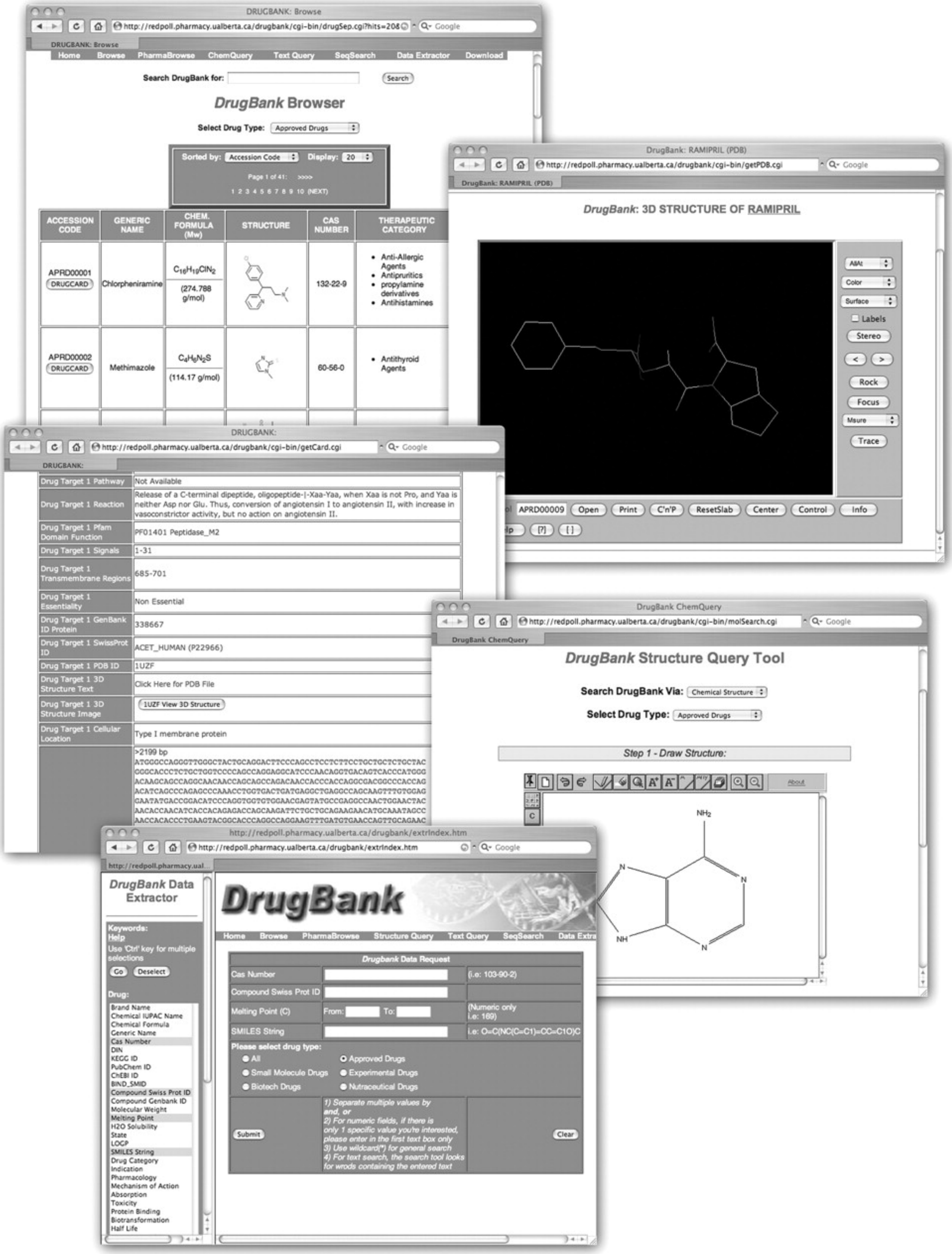
- The molecule must contain more than one type of atom;
- Be non-redundant;
- Have a known chemical structure and,
- Be identified as a drug or drug-like molecule by at least one reputable data source.

DrugBank is divided into four major categories:

1. FDA-approved small molecule drugs;
2. FDA-approved biotech (protein/peptide) drugs;
3. Nutraceuticals or micronutrients such as vitamins and metabolites and,
4. Experimental drugs, including unapproved drugs, de-listed drugs, elicit drugs, enzyme inhibitors and potential toxins.

These four categories can also be bundled into two categories, including FDA drugs (approved) and all compounds (Experimental, FDA and nutraceuticals) (Wishart *et al.*, 2006).

DrugBank is a fully searchable web-enabled resource with many built-in tools and features for viewing, sorting and extracting drug or drug target data. Detailed instruction on where to locate and how to use these browsing/search tools are provided on the DrugBank homepage. Like any web-enabled database, DrugBank supports standard text queries. It also offers general database browsing using the "Browse" and "PharmaBrowse" buttons located at the top of each DrugBank page. To facilitate more general browsing, DrugBank is divided into synoptic summary tables, which is again linked to more detailed "DrugCards" (analogous to the successful GeneCard concept). All of DrugBank's summary tables can be rapidly browsed, sorted or reformatted, similar to the way PubMed abstract can be viewed. By clicking on the DrugCard button it open a webpage describing the drug of interest in more detail. Each DrugCard entry contain > 80 data fields with half of the information devoted to drug/chemical data and the other half devoted to drug target or protein data (some of these fields can be seen as screen shots in Figure 1.7and a whole list can be seen in Table 1.1). In addition to providing comprehensive numeric, sequence and textual data, each DrugCard also contains hyperlinks to other



DRUGBANK: Browse
http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/drugSep.cgi?hits=208
Google
Home Browse PharmaBrowse ChemQuery Text Query SeqSearch Data Extractor Download
Search
DrugBank: RAMIPRIL (PDB)
http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/getPDB.cgi
AllAt Color Surface Labels Stereo < > Rock Focus Mture Trace
APRD00009 Open Print C'n'P ResetSlab Center Control Info
DRUGBANK:
http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/getCard.cgi
DrugBank ChemQuery
http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/molSearch.cgi
About
NH2
N
N
NH
N
http://redpoll.pharmacy.ualberta.ca/drugbank/extrIndex.htm
DrugBank
Home Browse PharmaBrowse Structure Query Text Query SeqSearch Data Extra
Go Deselect
Submit Clear

Search DrugBank for:

DrugBank Browser

Select Drug Type: Approved Drugs

Sorted by: Accession Code Display: 20

Page 1 of 41: >>>>

1 2 3 4 5 6 7 8 9 10 (NEXT)

ACCESSION CODE	GENERIC NAME	CHEM. FORMULA (Mw)	STRUCTURE	CAS NUMBER	THERAPEUTIC CATEGORY
APRD00001 DRUGCARD	Chlorpheniramine	$C_{16}H_{19}ClN_2$ (274.788 g/mol)		132-22-9	Anti-Allergic Agents Antipruritics propylamine derivatives Antihistamines
APRD00002 DRUGCARD	Methimazole	$C_4H_6N_2S$ (114.17 g/mol)		60-56-0	Antithyroid Agents

DrugBank: 3D STRUCTURE OF RAMIPRIL

Drug Target 1 Pathway	Not Available
Drug Target 1 Reaction	Release of a C-terminal dipeptide, oligopeptide-|-Xaa-Yaa, when Xaa is not Pro, and Yaa is neither Asp nor Glu. Thus, conversion of angiotensin I to angiotensin II, with increase in vasoconstrictor activity, but no action on angiotensin II.
Drug Target 1 Pfam Domain Function	PF01401 Peptidase_M2
Drug Target 1 Signals	1-31
Drug Target 1 Transmembrane Regions	685-701
Drug Target 1 Essentiality	Non Essential
Drug Target 1 GenBank ID Protein	338667
Drug Target 1 SwissProt ID	ACET_HUMAN (P22966)
Drug Target 1 PDB ID	1UZF
Drug Target 1 3D Structure Text	Click Here for PDB File
Drug Target 1 3D Structure Image	1UZF View 3D Structure
Drug Target 1 Cellular Location	Type I membrane protein
	>2199 bp ATGGGGCCAGGGTTGGGCTACTGCAGGACTTCCCAGCCTCCTCTTCCTGCTGCTCTGCTAC GGGCACCCTCTGCTGGTCCCCAGCCAGGAGGCATCCCAACAGGTGACAGTCACCCATGGG ACAAGCAGCCAGGCAACAACCAGCAGCCAGACAACCACCCACCAGGCGACGGCCCACCAG ACATCAGCCCAGAGCCCAAACCTGGTGACTGATGAGGCTGAGGCCAGCAAGTTTGTGGAG GAATATGACCGGACATCCCAGGTGGTGTGGAACGAGTATGCCGAGGCCAACTGGAACTAC AACACCAACATCACCACAGAGACCAGCAAGATTCTGCTGCAGAAGAACATGCAAATAGCC AACCACACCCTGAAGTACGGCACCCAGGCCAGGAAGTTTGATGTGAACCAGTTGCAGAAC

DrugBank Structure Query Tool

Search DrugBank Via: Chemical Structure

Select Drug Type: Approved Drugs

Step 1 - Draw Structure:

DrugBank Data Extractor

Keywords:
Help
Use 'Ctrl' key for multiple selections

Drug:
Brand Name
Chemical IUPAC Name
Chemical Formula
Generic Name
Cas Number
DIN
KEGG ID
PubChem ID
ChEBI ID
BIND_SMID
Compound Swiss Prot ID
Compound Genbank ID
Molecular Weight
Melting Point
H2O Solubility
State
LOGP
SMILES String
Drug Category
Indication
Pharmacology
Mechanism of Action
Absorption
Toxicity
Protein Binding
Biotransformation
Half Life

Drugbank Data Request		
Cas Number		(i.e: 103-90-2)
Compound Swiss Prot ID		
Melting Point (C)	From: To:	(Numeric only i.e: 169)
SMILES String		i.e: O=C(NC(C=C1)=CC=C1O)C

Please select drug type:
● All ⊙ Approved Drugs
● Small Molecule Drugs ● Experimental Drugs
● Biotech Drugs ● Nutraceutical Drugs

1) Separate multiple values by and, or
2) For numeric fields, if there is only 1 specific value you're interested, please enter in the first text box only
3) Use wildcard(*) for general search
4) For text search, the search tool looks for wrods containing the entered text

Figure 1.7: A screenshot montage of the DrugBank Database.

Showing several possible views of information describing the drug Ramipril. Not all the fields are shown.

From: Wishart *et al.* (2006).

Table 1.1: Summary of the data fields or data types in each DrugCard. From: Wishart *et al.* (2006).

Drug or compound information	Drug target or receptor information
Generic name	Target name
Brand name(s)/synonyms	Target synonyms
IUPAC name	Target protein sequence
Chemical structure/sequence	Target no. of residues
Chemical formula	Target molecular weight
PubChem/KEGG/ChEBI links	Target pI
Swiss-Prot/GenBank links	Target gene ontology
FDA/MSDS/RxList links	Target general function
Molecular weight	Target specific function
Melting point	Target pathways
Water solubility	Target reaction
pKa or pI	Target Pfam domains
LogP or hydrophobicity	Target signal sequences
NMR/MS spectra	Target transmembrane regions
MOL/SDF/PDF text files	Target essentiality
MOL/PDB image files	Target GenBank protein ID
SMILES string	Target Swiss-Prot ID
Indication	Target PDB ID
Pharmacology	Target cellular location
Mechanism of action	Target DNA sequence
Biotransformation/absorption	Target chromosome location
Patient/physician information	Target locus
Metabolizing enzymes	Target SNPs/mutations

databases, abstracts, digital images and interactive applets for viewing molecular structures (as shown in Figure 1.7) (Wishart *et al.*, 2006).

1.4.5 TDR Targets Database

The TDR Targets project seeks to use the availability of diverse datasets to ease the identification and prioritization of drugs and drug targets in neglected disease pathogens. This database functions as a website, where researchers can look for information on targets of interest, as well as a tool for prioritization of targets in whole genomes (TDR, 2011). Researchers can use the TDRtargets database as a tool to quickly prioritize genes of interest by running simple queries, like looking for small enzymes, or proteins with high quality structural models; assigning numerical weights to each query, in the history page; and combining these results to produce a ranked list of candidate targets (TDR, 2011).

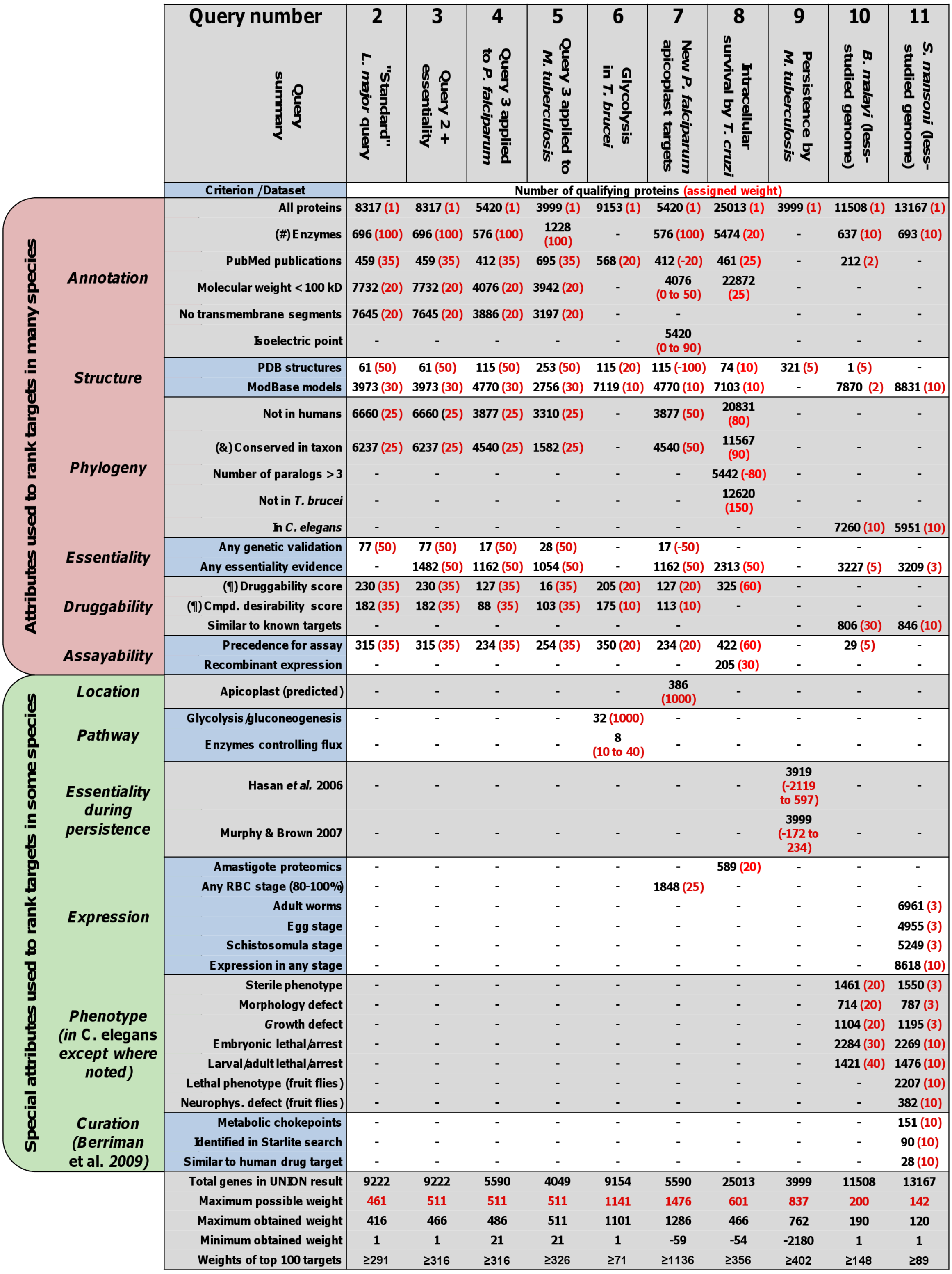
TDR contain mainly two datasets, namely, genomic dataset and chemical dataset. The genomic dataset contain the following information:

- Species;
- Basic genome annotation data;
- Functional annotation data;
- Structure data;
- Expression data;
- Antigenicity data;
- Phylogenetic distribution data;
- Essentiality data;
- Curated data on gene validation;
- Druggability data;
- Assayability data and,
- Bibliographic references.

The chemical dataset contains the following:

- Basic information;
- Chemical properties;
- Number of atoms;
- Pharmacological activity;
- Information source and,
- Structure based searches (TDR, 2011).

As a starting point in TDR, a basic set of criteria of general interest should be chosen to frame a “standard” query for identifying targets in e.g. *Leishmania major* (query 2, Figure 1.8). In compiling the basic set of criteria, TDR included most datasets that are commonly available for organisms with complete genomic information, so that the standard query could be easily applied to different pathogens. Queries 3, 4, and 5 of Figure 1.8 are examples of extending standard queries; queries 6, 7, 8, and 9 of Figure 1.8 are framed in a pathogen-specific manner to prioritize target proteins from a particular metabolic pathway, sub-cellular location, or lifecycle stage by uploading external datasets to TDRtargets.org; queries 10 and 11 of Figure 1.8 were based heavily on data obtained by



		Query number	2	3	4	5	6	7	8	9	10	11
		Query summary	"Standard" *L. major* query	Query 2 + essentiality	Query 3 applied to *P. falciparum*	Query 3 applied to *M. tuberculosis*	Glycolysis in *T. brucei*	New *P. falciparum* apicoplast targets	Intracellular survival by *T. cruzi*	Persistence by *M. tuberculosis*	*B. malayi* (less-studied genome)	*S. mansoni* (less-studied genome)
		Criterion /Dataset	Number of qualifying proteins (assigned weight)									
Attributes used to rank targets in many species	Annotation	All proteins	8317 (1)	8317 (1)	5420 (1)	3999 (1)	9153 (1)	5420 (1)	25013 (1)	3999 (1)	11508 (1)	13167 (1)
		(#) Enzymes	696 (100)	696 (100)	576 (100)	1228 (100)	-	576 (100)	5474 (20)	-	637 (10)	693 (10)
		PubMed publications	459 (35)	459 (35)	412 (35)	695 (35)	568 (20)	412 (-20)	461 (25)	-	212 (2)	-
		Molecular weight < 100 kD	7732 (20)	7732 (20)	4076 (20)	3942 (20)	-	4076 (0 to 50)	22872 (25)	-	-	-
		No transmembrane segments	7645 (20)	7645 (20)	3886 (20)	3197 (20)	-	-		-	-	-
		Isoelectric point	-	-	-	-	-	5420 (0 to 90)		-	-	-
	Structure	PDB structures	61 (50)	61 (50)	115 (50)	253 (50)	115 (20)	115 (-100)	74 (10)	321 (5)	1 (5)	-
		ModBase models	3973 (30)	3973 (30)	4770 (30)	2756 (30)	7119 (10)	4770 (10)	7103 (10)	-	7870 (2)	8831 (10)
	Phylogeny	Not in humans	6660 (25)	6660 (25)	3877 (25)	3310 (25)	-	3877 (50)	20831 (80)	-	-	-
		(&) Conserved in taxon	6237 (25)	6237 (25)	4540 (25)	1582 (25)	-	4540 (50)	11567 (90)	-	-	-
		Number of paralogs > 3	-	-	-	-	-	-	5442 (-80)	-	-	-
		Not in *T. brucei*	-	-	-	-	-	-	12620 (150)	-	-	-
		In *C. elegans*	-	-	-	-	-	-	-	-	7260 (10)	5951 (10)
	Essentiality	Any genetic validation	77 (50)	77 (50)	17 (50)	28 (50)	-	17 (-50)	-	-	-	-
		Any essentiality evidence	-	1482 (50)	1162 (50)	1054 (50)	-	1162 (50)	2313 (50)	-	3227 (5)	3209 (3)
	Druggability	(¶) Druggability score	230 (35)	230 (35)	127 (35)	16 (35)	205 (20)	127 (20)	325 (60)	-	-	-
		(¶) Cmpd. desirability score	182 (35)	182 (35)	88 (35)	103 (35)	175 (10)	113 (10)	-	-	-	-
		Similar to known targets	-	-	-	-	-	-		-	806 (30)	846 (10)
	Assayability	Precedence for assay	315 (35)	315 (35)	234 (35)	254 (35)	350 (20)	234 (20)	422 (60)	-	29 (5)	-
		Recombinant expression	-	-	-	-	-	-	205 (30)	-	-	-
Special attributes used to rank targets in some species	Location	Apicoplast (predicted)	-	-	-	-	-	386 (1000)	-	-	-	-
	Pathway	Glycolysis /gluconeogenesis	-	-	-	-	32 (1000)	-	-	-	-	-
		Enzymes controlling flux	-	-	-	-	8 (10 to 40)	-	-	-	-	-
	Essentiality during persistence	Hasan *et al.* 2006	-	-	-	-	-	-	-	3919 (-2119 to 597)	-	-
		Murphy & Brown 2007	-	-	-	-	-	-	-	3999 (-172 to 234)	-	-
	Expression	Amastigote proteomics	-	-	-	-	-	-	589 (20)	-	-	-
		Any RBC stage (80-100%)	-	-	-	-	-	1848 (25)	-	-	-	-
		Adult worms	-	-	-	-	-	-	-	-	-	6961 (3)
		Egg stage	-	-	-	-	-	-	-	-	-	4955 (3)
		Schistosomula stage	-	-	-	-	-	-	-	-	-	5249 (3)
		Expression in any stage	-	-	-	-	-	-	-	-	-	8618 (10)
	Phenotype (in *C. elegans* except where noted)	Sterile phenotype	-	-	-	-	-	-	-	-	1461 (20)	1550 (3)
		Morphology defect	-	-	-	-	-	-	-	-	714 (20)	787 (3)
		Growth defect	-	-	-	-	-	-	-	-	1104 (20)	1195 (3)
		Embryonic lethal/arrest	-	-	-	-	-	-	-	-	2284 (30)	2269 (10)
		Larval/adult lethal/arrest	-	-	-	-	-	-	-	-	1421 (40)	1476 (10)
		Lethal phenotype (fruit flies)	-	-	-	-	-	-	-	-	-	2207 (10)
		Neurophys. defect (fruit flies)	-	-	-	-	-	-	-	-	-	382 (10)
	Curation (Berriman et al. 2009)	Metabolic chokepoints	-	-	-	-	-	-	-	-	-	151 (10)
		Identified in Starlite search	-	-	-	-	-	-	-	-	-	90 (10)
		Similar to human drug target	-	-	-	-	-	-	-	-	-	28 (10)
		Total genes in UNION result	9222	9222	5590	4049	9154	5590	25013	3999	11508	13167
		Maximum possible weight	461	511	511	511	1141	1476	601	837	200	142
		Maximum obtained weight	416	466	486	511	1101	1286	466	762	190	120
		Minimum obtained weight	1	1	21	21	1	-59	-54	-2180	1	1
		Weights of top 100 targets	≥291	≥316	≥316	≥326	≥71	≥1136	≥356	≥402	≥148	≥89

Figure 1.8: A summary of the multi-parameter search queries presented by Crowther *et al.* (2010). Ten different queries (Queries 2–11) are listed as individual columns for which the criteria are shown on the left. For each criterion, the number of qualifying proteins from a given pathogen is shown in black and the associated weight is shown in red within parentheses. Symbols: (#) enzymes were selected by combining searches by EC number and by functional category, except for Queries 10 and 11, which were based only on EC number; (&) the conserved-in-taxon criterion refers to the presence of orthologs in *L. major, T. brucei*, and *T. cruzi, P. falciparum* and *P. vivax, M. tuberculosis* and *M. leprae*, and *L. major* and *T. cruzi*; (¶) druggability and compound desirability scores were queried using respective cutoff values of ⩾ 0.6 and > 0.3, ⩾ 0.4 and > 0.2, and ⩾ 0.5. From: Crowther *et al.* (2010).

manual curation of literature and homology/orthology analysis for protein-specific information. These illustrate how even incompletely annotated genomes are amenable to target identification (Crowther *et al.*, 2010).

1.4.6 Genomes2Drug

Genomes2Drug is a freely available web-based search engine that simultaneously searches each input protein sequence against the protein sequences of the human genome, the DrugBank dataset drug targets and the PDB protein structure database. A single FASTA formatted protein sequence or multiple sequences can be uploaded, and the results can be viewed in Excel (Toomey *et al.*, 2009).

Each E_{BLAST_P} value is derived from the optimal alignment across the genome using default settings of NCBI's freely available $BLAST_p$ algorithms. As the best alignment score is recorded for each input protein, it follows that a poor score indicates that there is no matching protein in the comparator set. Thus a large E_{BLAST_p}[query vs. human genome] value indicates that there is likely no match for that query protein in the human genome. Similarly, good sequence identity, with a small E_{BLAST_P}[query vs. PDB] value indicates that the query sequence has a close homologue in the PDB structural database (Toomey *et al.*, 2009). The <human expect> and <PDB expect> columns can be used individually to rank the whole input genome for proteins showing little homology to humans or good homology to a protein with a known 3D structure, respectively. The ratio of these expected values can also be used to rank the output list according to proteins that would be readily structurally modelled, while also showing little identity to any human proteins (Toomey *et al.*, 2009).

1.5 Conclusion

The methods described above all use different criteria and data to identify potential drug targets. As some of these techniques and data used are so divers it will be difficult to do comparisons between the different methods and databases; even the ranking techniques differ. But if as much as possible of these techniques are combined and all its criteria is taken into consideration it might be possible to get an approach that delivers a list of possible drug targets for an organism that have a minimal amount of false results.

To run thousands of proteins *in vitro* in a laboratory to identify potential targets for an organism can be quite a lengthy process. Thus, target identification and prioritization via *in silico* approaches is not meant as a replacement for *in vitro* work, but rather as a prelude to it, to restrict and help focus the amount of proteins that need to be tested from thousands to mere hundreds.

1.6 Problem Statement

Since most antimalarial drugs are showing a degree of resistance, due to an over-and misuse, new antimalarial drugs need to be discovered. The DISCOVERY Database aims to aid in this discovery by having a database filled with aspects that might influence the druggability of a protein and guide a scientist in choosing the right ligand for that protein. Over 200 genes of the *Plasmodium* genome might encode suitable drug targets. To sift through all these proteins in the laboratory, to find which one is the next first-line antimalarial drug, can be a lengthy process.

1.7 Specific Research Questions and Aims

The aim of this project is:

- To identify proteins that have a possibility of being druggable.

    - A workflow system will be developed to identify important drug target criteria for the *Plasmodium falciparum* (*P. falciparum*) proteins via DISCOVERY.
    - All the known *P. falciparum* proteins will be run through the DISCOVERY system, via a Taverna workflow, and scored at the end.
    - Weighing and adding all the scores for a particular protein will compose an aggregate score.
    - Proteins with a high score will be identified as being highly druggable, compared to proteins with a low score.
    - Four case studies will aid as examples as to how good the workflow and scoring system worked.

Chapter 2

In silico identification of drug targets using Taverna and DISCOVERY

2.1 Introduction

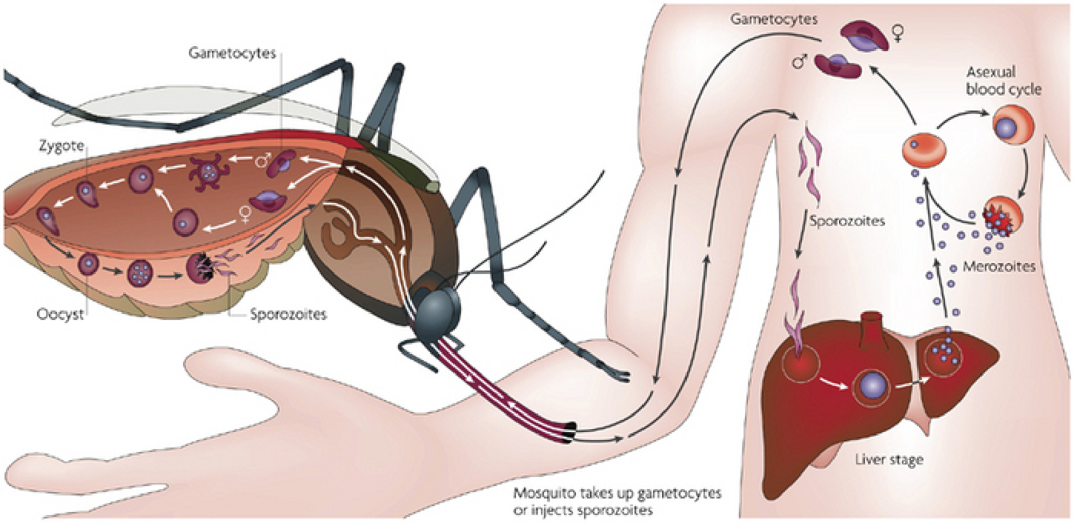
Parasitic diseases continue to take a huge toll on the human health, particularly in tropical areas. These diseases can include malaria, caused by the *Plasmodium* spp.; leishmaniasis, caused by *Leishmania* spp.; African trypanosomiasis, caused by *Trypanosoma brucei gambiense* and *T.b. rhodesiense*; chagas disease, caused by *Trypanosoma cruzi*; schistosomiasis, caused by *Schistosoma mansoni*, *S. haematobium*, and *S. japonica*; lymphatic filariases, that is caused by *Brugia malayi* and *Wuchereria bancrofti*; and onchocerciasis, caused by *Onchocerca volvulus* (Pink *et al.*, 2005).

2.1.1 Malaria

The *Plasmodium* spp. is an obligate intracellular protozoan parasite of humans and other animals where the transmission occurs via the *Anopheles* mosquito vector (Aurrecoechea *et al.*, 2009; Gardner *et al.*, 2002; Li *et al.*, 2011).

Malaria is a life threatening disease, affecting half a billion people in 109 countries and territories, from South America to the Indian peninsula (Aurrecoechea *et al.*, 2009; Fatumo *et al.*, 2011; Li *et al.*, 2011; Travassos and Laufer, 2009). Malaria causes approximately 2 million deaths annually (Aurrecoechea *et al.*, 2009; Gardner *et al.*, 2002; Li *et al.*, 2011; Schunk *et al.*, 2006; Travassos and Laufer, 2009), many of which are children under the age of 5 years (Aurrecoechea *et al.*, 2009; Birkholtz *et al.*, 2006; Gardner *et al.*, 2002; Joubert *et al.*, 2009).

Four of the most important malaria species that infect humans, via the mosquito vector, include *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale* and *Plasmodium malariae*. All of them belong to the phylum *Apicomplexa* that also includes other parasitic protozoa (Birkholtz *et al.*, 2006; Gardner *et al.*, 2002). *P. falciparum* is one of the deadliest *Plasmodium* species, causing malaria in humans (Aurrecoechea *et al.*, 2009). It causes 80% of the world's infections and 90% of the deaths (Joubert *et al.*, 2009). The genome sequence of *P. falciparum* was released in November 2002. It has a 23 Mb nuclear genome, consisting of 14 chromosomes and encoding $5,300$ genes (Gardner *et al.*, 2002). Compared to other genomes of free-living eukaryotic microbes, *P. falciparum* encodes fewer enzymes and transporters (these can be undiscovered enzymes, not necessarily missing enzymes (Fatumo *et al.*, 2011)), but a large proportion of its genome is dedicated to immune evasion and host-parasite interaction (Gardner *et al.*, 2002). Through the completion of this genome the functional genomics could be observed (Fatumo *et al.*, 2011). *P. vivax* is responsible for more than $25 - 40\%$ of the approximately



Nature Reviews | Genetics

Figure 2.1: Plasmodium life cycle.
The Plasmodium life cycle is complicated and takes it through multiple cell types during which the parasite undergoes multiple developmental changes. The parasite alternates between the sexual stage in the insect vector and the asexual stage in the vertebrate host.
Source: http://www.nature.com/scitable/content/plasmodium-falciparum-life-cycle-14465535.

525 million cases of malaria. It is the most important malaria parasite that infects people outside of Africa, and is normally seen in Asia and America. *P. vivax* consist of a 26.8 Mb nuclear genome, containing $\sim 5,433$ genes and is spread over 14 chromosomes. *P. vivax* can cause the patient to have a relapse after a few of months, or even years (Carlton *et al.*, 2008; Pain and Hertz-Fowler, 2009).

Other *Plasmodium* species that are used mainly for research, as these are malaria parasites that infect rodents, are *P. berghei* and *P. chabaudi*. Both have 14 chromosomes. *P. berghei* has a genome size of 18 Mb and *P. chabaudi* has a genome size of 17 Mb (Hall *et al.*, 2005). The sequencing of these genomes has shown that there are $\sim 4,500$ genes conserved in all *Plasmodium* spp. and are located in central regions of the 14 chromosomes (Hall *et al.*, 2005).

2.1.1.1 Lifecycle

The life cycle of this parasite is complicated and takes it through multiple cell types during which the parasite undergoes multiple developmental changes (see Figure 2.1) (Aurrecoechea *et al.*, 2009). The parasite alternates between two stages: the sexual stage in the insect vector, the *Anopheles* mosquito, and the asexual stage in the vertebrate host, human or any other mammal (Pain and Hertz-Fowler, 2009). In the mosquito (sexual stage) the gametocytes are ingested when the mosquito feeds on an infected person. The gametocytes then differentiate into male and female gametes. The gametes fuse and form a zygote in the gut of the mosquito; it differentiates further, and forms an ookinete. The ookinete penetrates the gut lining, producing an oocyst. When the oocyst ruptures it releases sporozoites that migrate to the salivary gland, ready for infecting a host (Figure 2.1) (Menard, 2005; Pain and Hertz-Fowler, 2009). In the human (asexual stage) sporozoites enter the liver. In the liver they multiply asexually and differentiate into merozoites. The merozoites escape into the blood stream where it infects red blood cells (RBC). In the RBC it differentiates and forms trophozoites. The trophozoites multiply asexually again and re-differentiate into merozoites where it ruptures the host

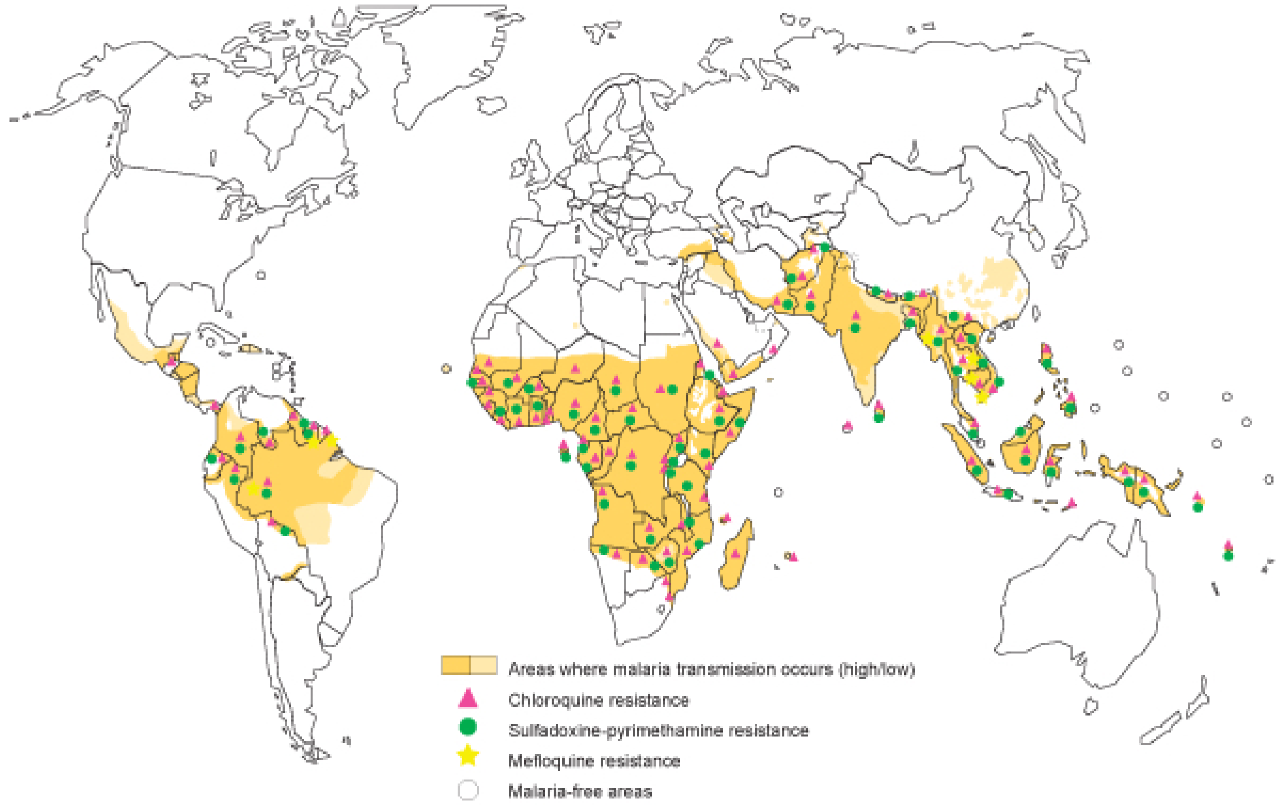

Areas where malaria transmission occurs (high/low)
▲ Chloroquine resistance
● Sulfadoxine-pyrimethamine resistance
★ Mefloquine resistance
○ Malaria-free areas

Figure 2.2: Drug resistance to *P. falciparum* from studies, up to 2004.
The yellow-highlighted areas indicate regions of malaria transmission, the darker the yellow the higher the rate. Chloroquine, sulfadoxine-pyrimethamine, and mefloquine resistance areas are marked. Source: http://www.rbm.who.int/wmr2005/html/map5.htm.

RBC to infect new ones. At any time, between the merozoite stage and trophozoite stage, a gametocyte can form; ready to be ingested by a mosquito (Figure 2.1) (Menard, 2005).

2.1.1.2 Drug resistance

Drug resistance to most classes of antimalarial agents has been observed, including artemisinin (Gardiner *et al.*, 2009; Joubert *et al.*, 2009; Mpangase *et al.*, 2013; Travassos and Laufer, 2009). This is caused by the over and misuse of antimalarial drugs and is spreading at an alarming rate throughout the endemic areas (see Figure 2.2) (Joubert *et al.*, 2009). Attempts to achieve progress in malaria prophylaxis include failure in vaccine development, the withdrawal of some insecticides because of toxic and negative environmental impact, spread of mosquito resistance to insecticides and of resistance of the parasite to the few developed drugs available (Birkholtz *et al.*, 2006).

Drug resistance to sulfadoxine-pyrimethamine (SP) and chloroquine are frequently observed in some areas and fuel the on going burden of malaria (Travassos and Laufer, 2009). Due to this the World Health Organization (WHO) has recommended that the first line therapy to be combination treatments that include artemisini based combination therapy (ACTs) (Travassos and Laufer, 2009). But new drugs, even if used in combination are not resistant to resistance (Travassos and Laufer, 2009) and some ACTs have already started to show some resistance, especially in Southeast Asia (Mpangase *et al.*, 2013).

Resistance is spreading rapidly for chloroquine (Birkholtz *et al.*, 2006; Lopes *et al.*, 2002). Chloroquine acts by interfering with the detoxification of haematin (Travassos and Laufer, 2009). Chloroquine resistance has been identified as due to a mutation in the *P. falciparum* chloroquine resistance transporter (PfCRT). In the mutated form it is associated with decreased chloroquine accumulation in the food vacuole of the parasite. PfCRT is not the only gene that mediates chloroquine resistance; *P. falciparum* multi-drug resistance gene (PfMDR) also plays a role in chloroquine resistance. The genetic background of the strains can also influence the degree of resistance conferred by these mutations (Schunk *et al.*, 2006; Travassos and Laufer, 2009). The resistance to chloroquine has developed independently in Asia, Papua New Guinea and South America and the extensive use of chloroquine increases the prevalence of the PfCRT mutated gene (Travassos and Laufer, 2009). Even though chloroquine is not used anymore in the treatment of *P. falciparum* it is still widely used to treat *P. vivax* and *P. ovale*, that are less severe forms of malaria and cause recurrent infections (Travassos and Laufer, 2009). Evidence from genetic crosses indicated that the PfMDR gene does not only play a role in chloroquine resistance, but can also adjust sensitivity to mefloquine, quinine, halofantrine and structurally different artemisinin in *P. falciparum* (Travassos and Laufer, 2009).

Due to chloroquine resistance the first-line antimalarial drug was changed to other antimalarial drugs, including SP. SP is composed of two drugs; pyrimethamine, a diamino-pyrimidine and antifolate, and sulfonamides, as sulfadoxine. These drugs act on enzymes in the folate synthesis pathway. If both SP drugs are used they work synergistically. But, resistance to SP has also been observed and has spread faster than it had for chloroquine. The resistance of *P. falciparum* to SP is caused by mutations of the *P. falciparum* dihydropteroate synthase (PfDHPS) and *P. falciparum* dihydrofolate reductase (PfDHFR) genes (Travassos and Laufer, 2009). SP was, until relatively recently, the drug of choice for uncomplicated malaria and for the intermittent preventative treatment of malaria for the vulnerable group that includes pregnant women, infants and children (Travassos and Laufer, 2009).

ACTs have been introduced, first in Asia and then in Africa and South America. The mechanism of action has not been fully explained, but it may inhibit sarco/endoplastic reticulum calcium ATPase that can be found in *Plasmodium* parasites. Artemisinins act rapidly and kill the *Plasmodium* parasite throughout the asexual blood stages; it can also decrease transmission because it acts on the gametocytes. Due to short therapy duration of 5 to 7 days artesunate cannot be administered alone due to a high risk of recurring infection. When artemisinin is combined with a longer acting drug it can be administered as an ACT over a couple of days (Travassos and Laufer, 2009).

2.1.1.3 Drug discovery

The drugs that are used for malaria are far from perfect as many were introduced decades ago (Pink *et al.*, 2005) and most of them are showing a high rate of resistance outbreaks, especially in *P. falciparum* (Rosenthal, 2003). Due to this there is a constant need for the discovery of new antimalarials as well as drug targets (de Beer *et al.*, 2009; Gardiner *et al.*, 2009; Joubert *et al.*, 2009; Orti *et al.*, 2009; Rosenthal, 2003).

The complete *P. falciparum* genome sequence, and more recently other malaria species, has allowed the identification of new molecular targets within the parasite that may be exploited for drug design and vaccine purposes (Gardner *et al.*, 2002; Orti *et al.*, 2009), but this requires careful and rational selection of suitable proteins and corresponding ligands. This is the primary goal of most *Plasmodium* genome projects (Travassos and Laufer, 2009). The trend in the search for new antimalarial drugs was to use target-based drug discovery, but the disadvantage is that predicting which targets will be essential for parasite growth in mammalian hosts is often difficult; well validated targets may not necessary yield new classes of compounds and resistance may emerge more quickly for the compounds that interact only with a singular target (Dharia *et al.*, 2010). This has led to recent drug screening effort having shifted to cell-based assays and yielding thousands of potential antimalarial compounds. But a challenge remains, namely, to discover the mode of action of these compounds (Dharia *et al.*, 2010).

The approaches must take into account specific concerns, like the requirement for very inexpensive and simple-to-use new therapies and the need to limit the cost of drug discovery (Rosenthal, 2003). Antimalarial drug development can follow several strategies, ranging from minor modifications of existing agents to the design of novel agents that act against new targets. Increasingly, available agents are being combined to improve antimalarial regimens (Rosenthal, 2003).

Limitations

Antimalarial drug development has the same limitations, as many other drug development programs in that new agents must demonstrate efficacy, be safe and have additional properties important for the specific disease indication (Rosenthal, 2003). Market forces are insufficient to drive the discovery and development of antimalarial drugs. Between 1975 and 1999 more than $1,300$ drugs were introduced for all indications, of these only 13 were for tropical diseases (Pink *et al.*, 2005). In 2000 approximately 0.1% of global investment in health research was devoted to drug discovery for selected tropical diseases; which included malaria, leishmaniasis and trypanosomiasis and tuberculosis (Pink *et al.*, 2005). This is because most of these cases are in developing countries with limited resources. It is generally agreed that new antimalarials should be dosed orally and be effective with single-daily dosing, and that curative regimens should be short, ideally $1-3$ days in length. Thus the critical consideration in antimalarial drug development is an economic one (Rosenthal, 2003).

It has been estimated that the cost of bringing a new antimalarial drug to the market is approximately US$300 million, compared to US$500 million for a new a drug for all indications. The risk of failure in Phase II clinical trails is estimated to be 50% for new antimalarials; this is lower than the corresponding risk for a non-infectious disease (Pink *et al.*, 2005). Financial constraints are relevant for 2 reasons:

1. For the drug to be widely useful it needs to be very inexpensive so that it can be routinely available to populations in need in developing countries. A price of $1 per treatment is probably unacceptable in many regions. And,
2. Since malaria markets are primarily in developing countries, marketing opportunities have generally been considered to be limited, and so investment in antimalarial drug discovery and development has been small. Thus, drug discovery directed against malaria is particularly reliant upon shortcuts that may obviate excessive cost (Rosenthal, 2003).

Approaches to drug discovery

Different basic approaches to drug discovery for malaria can be classified as short-to-medium term or long term. Short-to-medium term is new drugs based on the exploitation of existing compounds, or compound classes while long term is new drugs that require the discovery of new chemical classes (Pink *et al.*, 2005).

Optimization of therapy with existing agents:

Optimization therapy with existing agents may include new dosing regimens or formulations. Combination therapies, including newer agents (e.g. artemisinin derivatives, atovaquone) and new combinations of older agents (e.g. amodiaquine/sulfadoxine/pyrimethamine, chloroguanil/dapsone), were under study as first-line therapies for Africa and other areas with widespread drug resistance (Rosenthal, 2003). Combination of existing drugs offers possibilities of synergy, reduced toxicity, shorter treatment regimens and slowing the development of resistance (Pink *et al.*, 2005; Rosenthal, 2003).

Drug resistance reversers:

The combination of previously effective agents with compounds that reverse parasite resistance to these agents offers another possible approach for an antimalarial drug (Pink *et al.*, 2005; Rosenthal, 2003). Many drugs have been shown to reverse the resistance of *P. falciparum* to chloroquine *in vitro*. This has included the antihypertensive, verapamil, and the antidepressant, desipramine (Rosenthal, 2003). In many cases, unacceptably high concentrations of the resistance reversers are needed for the desires effects, but by combining two or more of these agents at pharmacological concentrations it may provide a clinically relevant resistance reversal (Rosenthal, 2003).

Development of analogs of existing agents:

In the medium term, analogues of existing antimalarial drugs might prove effective by improving upon existing antimalarials (Pink *et al.*, 2005; Rosenthal, 2003). This approach does not require knowledge of the mechanism of action or the biological target of the parent compound. Indeed, this approach was responsible for the development of many existing antimalarials. For example, chloroquine, primaquine and mefloquine were discovered through chemical strategies to improve upon quinine (Rosenthal, 2003). Novel analogues of pyrimethamine are being designed to overcome drug resistance resulting from mutations in dihydrofolate reductase (Pink *et al.*, 2005).

Compounds active against other diseases:

Another approach is to identify agents that are developed or marketed as treatments for other diseases. This is an attractive short-term strategy that offers savings in development time and expenses (Pink *et al.*, 2005). These compounds might act against orthologs of their targets in other systems or by different mechanisms against malaria parasites. The advantage of these compounds is that they have already been developed for a human indication, thus will be quite inexpensive to develop as antimalarials (Rosenthal, 2003). Many antiparasitic drugs first entered development for other indications, e.g. DB289 was initially used to treat *Pneumastic* pneumonia, but is now in clinical trials as a potential oral treatment for malaria and early stage African trypanosomiasis (Pink *et al.*, 2005).

Focused sample collections:

An alternative approach to screening large libraries of compounds against whole parasites is to screen focused sample collections. In this approach the emphasis is on identifying compounds with either defined biological effects against related parasites, or biochemical activity against an isoenzyme, or receptors related to the known molecular targets of other organisms (Pink *et al.*, 2005). Work on parasite genome sequences coupled with biochemical investigations has pinpointed enzymes; like protein farnesyl transferases, cysteine proteases, histone deacetylases and fattyacyl synthases as potential drug targets for malaria and other parasites. Opportunities to access these compounds and evaluate them for their antiparasitic activity need to be exploited, as the compounds can be a valuable source of new leads (Pink *et al.*, 2005).

De novo discovery

Whole parasite assays:

Longer-term strategies aim to discover novel active substances unrelated to known drugs. For parasitic diseases there is an alternative approach based on screening and analysing compounds for their activity against whole parasites. Screening diverse compound collections on whole parasites *in* *vitro* has been steadily declining during the past two decades, but it is now undergoing a renaissance due to assay improvement (Pink *et al.*, 2005). This is especially true for test systems using *P. falciparum*. Screening now often relies on the use of parasites transfected with reporter genes to enable easy, rapid detection of antiparasitic activity (Pink *et al.*, 2005).

Molecular targets and HTS:

HTS against molecular targets has become a preferred method of early drug discovery for much of the biopharmaceutical industry, it has recently been used to a wide extent in the search for new drugs for neglected parasitic diseases, like malaria (Pink *et al.*, 2005).

Moving from hits to leads to drug discovery:

The same general pattern; from "hit" to "lead" to "drug candidate"; is followed for the discovery of antiparasitic and other drugs (Figure 2.3). Compounds are selected for improved efficacy and pharmaceutical properties by studies of analogues and iterative medicinal chemistry. The structure of the target molecule can be helpful in directing medicinal chemistry efforts. An advantage for the discovery of new antimalarial drugs is the existence of good, highly predictive *in vitro* and *in vivo* assays for activity; these studies often use the same organism to those that infect the human patient (Pink *et al.*, 2005).

Even though parasitic strains used in laboratory tests are the same, or similar, to those infecting humans there are certain cases where the differences are important. The standard animal models for malaria infections use *Plasmodium berghei*, *Plasmodium chabaudi*, *Plasmodium yoelii* or *Plasmodium vinckei* (*P. vinckei*), instead of the *Plasmodium* species that infect humans, like *P. falciparum*. The differences can be crucial when a molecular target-based drug discovery strategy is followed. For example cysteine protease of *P. vinckei* (vinckepains) differs from the protease of *P. falciparum* (falcipains), so that both types of protease had to be expressed and studied in a falcipain-based program. Genetically modified parasites in which the pathogen target replaces the model target gene would be one solution to the problem (Pink *et al.*, 2005).

2.1.2 Pipelines

Pipelining is a natural concept that is seen in everyday life, like an assembly line it is a technique implementation where multiple instructions are overlapped in execution (Prabhu, 2009). Consider the assembly of a car: assume that certain steps in the assembly line are to install the engine, the hood, and the wheels (in that order). The car can have only one of the three steps done at once. After the engine is installed it moves on to having its hood installed, leaving the engine installation facilities available for the next car. The first car then moves on to wheel installation, the second car to hood installation, and a third car begin to have its engine installed. If engine installation takes 20 minutes, hood installation takes 5 minutes, and wheel installation takes 10 minutes, then to finish all three cars, one at a time, will take 105 minutes (35×3). But by using an assembly line the total time will only be 75 minutes ($35 + 20 + 20$).

The computer pipeline is divided in stages. Each stage completes a part of an instruction in parallel. The stages are connected one to the next to form a pipe - instructions enter at one end, progress through the stages, and exit at the other end (Prabhu, 2009). Thus, a pipeline can be defined as a graph that describes the order of, and mutual relationships between, the analyses to be performed on an input dataset (Fiers *et al.*, 2008).

With the increasing number of bioinformatics databases and computational programs being made available as web services, a workflow system is an essential tool for e-Scientists if they are to take full

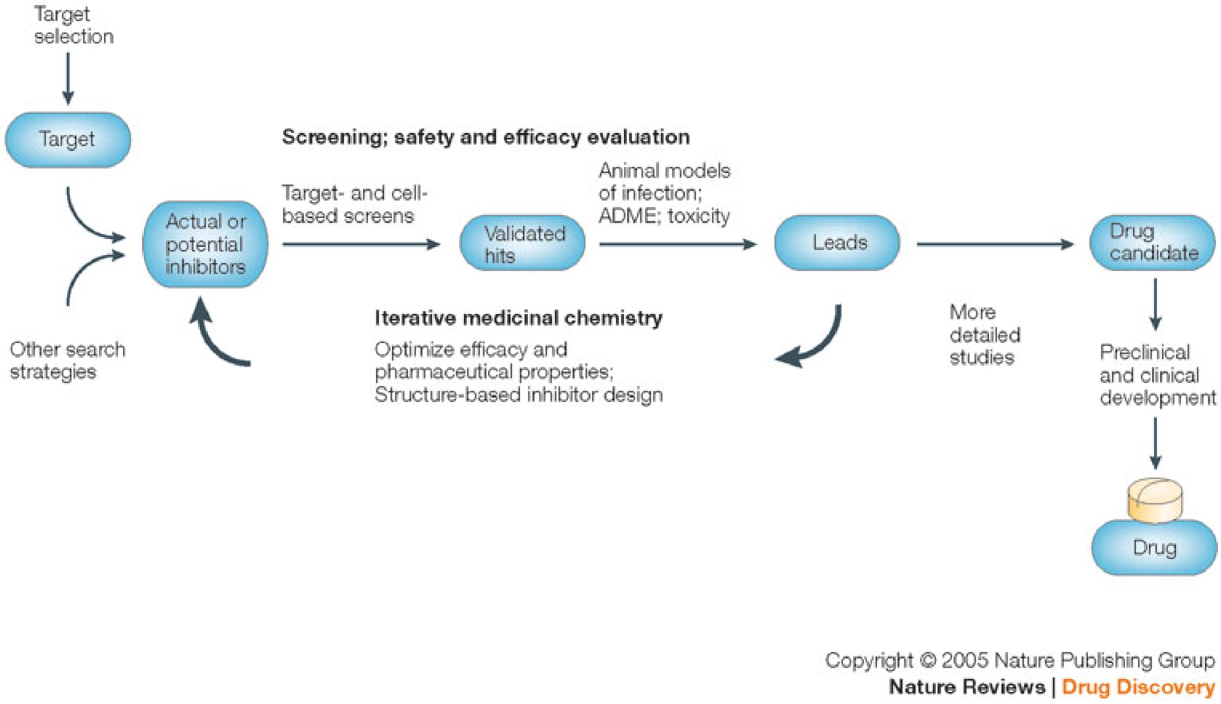

Target selection
Target
Other search strategies
Actual or potential inhibitors
Screening; safety and efficacy evaluation
Target- and cell-based screens
Validated hits
Animal models of infection; ADME; toxicity
Leads
More detailed studies
Drug candidate
Preclinical and clinical development
Drug
Iterative medicinal chemistry
Optimize efficacy and pharmaceutical properties; Structure-based inhibitor design
Copyright © 2005 Nature Publishing Group
Nature Reviews | Drug Discovery

Figure 2.3: Schematic illustrating the stages in drug discovery.

Drug discovery is a repetitive process that involves discrete stages. It normally begins with basic exploratory biology and biochemistry to identify molecular targets. In some cases compounds are tested, without knowledge of the target, for activity against the whole parasite. Compounds are assayed for activity against the target, if known, and for activity against the whole parasite. Compounds active against the whole parasite are defined as hits that can be considered for further testing in animal models of the disease. Other tests that monitor the compounds' pharmacokinetic properties are also initiated at this stage. Compounds that are active in the animal models and considered to be 'druggable' are defined as leads. Lead compounds generally require optimization for efficacy and good pharmaceutical properties. The process of optimization for pharmaceutical properties (adsorption, distribution, metabolism and excretion (ADME)) and lack of overt drug toxicity, as well as for efficacy against the target organism, is crucial. Once a compound reaches the stage at which it can be considered for testing in human patients, it is defined as a drug candidate. From there it enters the preclinical and then clinical studies of a typical drug development pathway.

From: Pink *et al.* (2005).

advantage of such resources (Oinn *et al.*, 2004). Even for small bioinformatics projects, with a few interdependent analyses, it is cumbersome to perform all operations manually and for larger projects this quickly becomes excessively complex (Fiers *et al.*, 2008).

2.1.2.1 Drawbacks

Pipelining does not decrease the time for individual instruction execution, but it increases the instruction throughput. The throughput of the instruction pipeline is determined by how often an instruction exits the pipeline (Prabhu, 2009). A major challenge for any pipeline system is to devise a fast and robust way to conduct data through a pipeline (Fiers *et al.*, 2008).

2.1.2.2 Taverna

The Taverna project was developed as an open source workflow tool enabling scientists to orchestrate bioinformatics web services and existing bioinformatics applications in workflows. The initial users of Taverna are bioinformaticians who can both develop and run workflows. It provides a graphical workbench tool for creating and running workflows that represent *in silico* bioinformatics experiments. In Taverna, a workflow is considered to be a graph of processors, each of which transforms a set of data inputs into a set of data outputs (Oinn *et al.*, 2004). It eases the use and integration of the growing number of molecular biology tools and databases available on the web (Hull *et al.*, 2006). It allows bioinformaticians, who are not necessarily expert in web services and programming, to construct workflows or pipelines of services to perform a range of different analyses, such as sequence analysis and genome annotation. These high-level workflows can integrate many different resources into a single analysis (Hull *et al.*, 2006).

A Taverna workflow consists mainly of three main entities:

1. **Processors:** A processor is a transformation that accepts a set of input data and produces a set of output data. Processors have a name within the workflow and a set of both input and output ports. During the execution of a workflow, each processor has a current execution status, which is one of initializing, waiting, running, complete, failed or aborted. A workflow can also possess input and output data entities:
   (a) A workflow input could be considered to be a source processor that executes instantaneously and makes the input value available on its virtual output port;
   (b) A workflow output can be considered as a sink processor that receives a value from its virtual input port but never actually executes. Both workflow sources and sinks can be annotated with metadata.
2. **Data links:** Data links mediate the flow of data between a data source and a data sink. The data source can be a processor output or a workflow input. The data sink can be a processor input port or a workflow output. Each data sink will receive the same value if there are multiple links from a data source. And,
3. **Coordination constraints:** A coordination constraint links two processors and controls their execution. This level of control is required when there is a process where the stages must execute in a certain order and yet there is no direct data dependency between them. For example, coordination constraints can be used to allow one processor to go from scheduled to running if another processor has status completed. In most cases, no concurrency constraints are required since data links will ensure that some processors stay in their waiting state until the data they require is available (Oinn *et al.*, 2004).

2.1.3 Web services

Web services are services that are made available from a business's web server for web users or other web-connected programs; it might include a combination of programming, data and human resources. It describes a standardized way of integrating web-based applications using the XML (eXtensible Markup Language), SOAP (Simple Object Access Protocol), WSDL (Web Services Description Language) and UDDI (Universal Description, Discovery and Integration) open standards over an Internet protocol backbone. It is used primarily as a means for businesses to communicate with each other and with clients. Web services allow organizations to communicate data without intimate knowledge of each other's IT systems behind the firewall and can range from such major services as storage management and customer relationship management down to much more limited services such as the furnishing of a stock quote and the checking of bids for an auction item (Beal, N/A; TechTarget, 2007).

Users can access some web services through a peer-to-peer arrangement rather than by going to a central server. Some services can communicate with other services and a class of software known as middleware generally enables this exchange of procedures and date. Unlike traditional client/server models, such as a web server/web page system, web services do not provide the user with a GUI (Graphic User Interface) instead it shares business logic, data and processes through a programmatic interface across a network. It is the applications that interact, not the users. Services previously possible only with the older standardized service known as Electronic Data Interchange increasingly are likely to become web services (Beal, N/A; TechTarget, 2007).

Developers can add the web service to a GUI, like a web page or an executable program, to offer specific functionality to users. Web services allow different applications from different sources to communicate with each other without time-consuming custom coding, and because all communication is in XML, web services are not tied to any one operating system or programming language. For example, Java can "talk" with Perl, Windows applications can "talk" with UNIX applications (Beal, N/A).

2.1.3.1 Wrappers

A wrapper is data that precedes or frames the main data or a program that sets up another program so that it can run successfully. It is intended to be an isolation point for the portion of code that:

1. De-serializes the operation and parameter data from a SOAP request;
2. Calls the application that implements the web service and,
3. Serializes the return from the application and builds the body portion of the SOAP response (IBM, 2013; TechTarget, 2005).

Wrappers can be used for different things, and include:

- On the Internet, "http://" and "ftp://" are sometimes described as wrappers for the Internet addresses or the Uniform Resource Locators (URL) that follow. A set of bracketing symbols, like "<" and ">" are sometimes referred to as wrappers;
- In programming, a wrapper is a program or script that sets the stage and makes it possible for another, more important program to run;
- In data transmission, a wrapper is the data that is put in front of, or around, a transmission that provides information about it and may also encapsulate it from view to anyone other than the intended recipient. A wrapper often consists of a header that precedes the encapsulated data and the trailer that follows it. And,

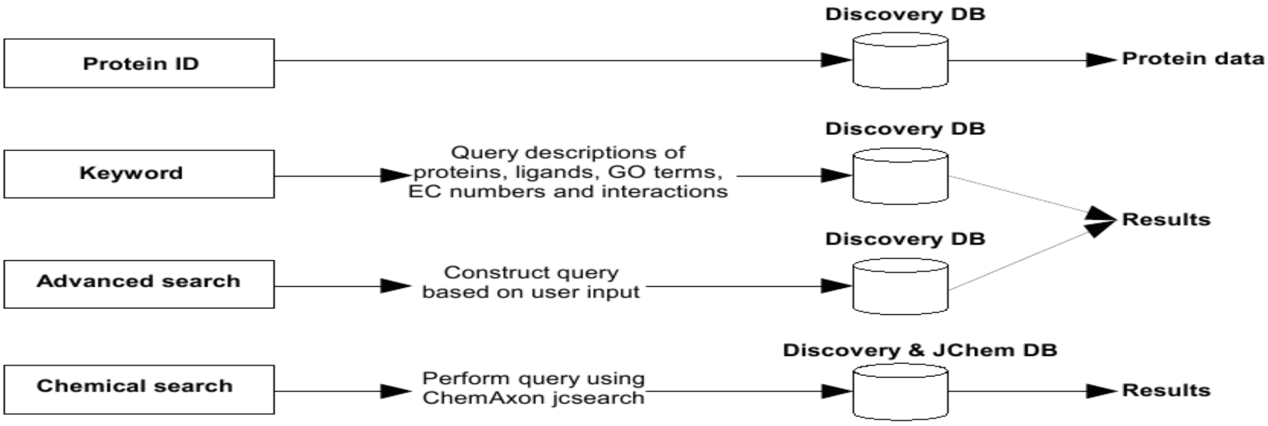

Protein ID
Discovery DB
Protein data
Keyword
Query descriptions of proteins, ligands, GO terms, EC numbers and interactions
Discovery DB
Results
Advanced search
Construct query based on user input
Discovery DB
Chemical search
Perform query using ChemAxon jcsearch
Discovery & JChem DB
Results

Figure 2.4: A representation of the different types of searches available in the Discovery system. Inserting a PlasmoDB/Ensembl identifier, UniProt accession number or the protein name in the basic search will result in the data on that specific protein. The advanced search function allows the user to filter proteins using different search criteria to select proteins matching the criteria. A list of matching proteins will be returned for the user to select the desired protein. When approaching the chemical search, the user may search by ligand keyword, or by drawing a chemical structure. Different methods are then available to search with the structure against the ligands.
From: Joubert *et al.* (2009).

- In database technology, a wrapper can be used to determine who has access to look at, or change the data that is wrapped (TechTarget, 2005).

Wrappers allow the application to be deployed as a web service while not requiring changes to handle request data in the SOAP format (IBM, 2013).

2.1.4 Malaria databases

There are quite a few databases available for users to use but none, or very few, incorporate all the known data available for malaria and compare it to the host and vector.

2.1.4.1 DISCOVERY

DISCOVERY is a web-based system that was developed in Java with NetBeans for the *in silico* selection of drug target proteins and lead compounds, and attempts to predict the interaction of ligands with proteins of interest. DISCOVERY attempts to associate chemical compound with malaria proteins by using sequence homology and also selective chemical similarity searches (Mpangase *et al.*, 2013). In DISCOVERY researchers can mine information on malaria proteins, predict ligands and compare human and mosquito host characteristics (Joubert *et al.*, 2009). Figure 2.4 illustrate the different types of searches in which a researcher can start to mine this information.

It includes protein sequences from *P. falciparum*, *P. vivax*, *P. chabaudi*, *P. berghei* and *P. yoelii* which were downloaded from PlasmoDB (version 8.1) as well as *Homo sapiens* and *Anopheles gambiae* which were downloaded from Ensembl (version 64) and VectorBase (version 3.7), respectively. Protein information includes sequences and annotations, obtained from PlasmoDB, Ensembl and VectorBase; functional predictions; gene ontology terms; orthology information; structural information; metabolic pathways; predicted putative protein-ligand interactions; druggability predictions and literature links.

The UniProt accessions were assigned to the proteins using an identifier-mapping file downloaded from UniProt.

The UniProt accessions were used to assign Gene Ontology (GO) annotations from UniProt-GOA database (October 2012). Each GO term was linked to the Gene Ontology database and the AmiGO visualization tool is used to view the GO terms (Mpangase *et al.*, 2013). To detect orthologous genes between the different species, ortholog clusters were generated with OrthoMCL using mostly default parameters with a percentage match cut-off of 50% (Mpangase *et al.*, 2013). Protein families, domains and functional sites were identified using InterProScan, which uses an ensemble of different methods to automatically annotate the function and structure of protein sequences (Mpangase *et al.*, 2013). To identify possible structures of related proteins, a sequence similarity search against the PDB database was carried out using NCBI BLAST, with an E-value cut-off value of 10^{-6}. Predicted 3D protein structures for *P. falciparum* and *H. sapiens* were downloaded from the ModBase database, which contains automatically generated homology models as part of the MODELER project (Mpangase *et al.*, 2013). Pathway information was obtained from KEGG PATHWAYS, MPMP (which is the primary malaria metabolic pathways site) and Reactome. Proteins where assigned to pathway information. This was done by creating links to the pathways in the different databases where they are involved in (Mpangase *et al.*, 2013). The Enzyme Commission (EC) numbers were assigned to the proteins using enzyme data downloaded from the ENZYME database hosted by ExPASy. For *Plasmodium* species, the EC numbers were obtained from PlasmoDB. The EC numbers were subsequently used to link proteins to the enzyme data in BRENDA and KEGG ENZYME (Mpangase *et al.*, 2013). The ChEMBL group kindly provided protein sequences from the DrugEBIlity database, which contains druggability predictions on protein structures from PDB. A sequence similarity search against the protein sequences from DrugEBIlity was carried out using NCBI BLAST, with an E-value cut-off of 10^{-6}, in order to assess the druggability of the matching proteins in DISCOVERY2 as well as closely-related proteins. The hits were linked to the DrugEBIlity database for more detailed information on the druggability calculations (Mpangase *et al.*, 2013). Data for protein-ligand interactions were collected from the ChEMBL database. Performing a BLAST search against the ChEMBL protein targets generated predictions. Matching functional domains that were generated through the InterProScan program was then found (Mpangase *et al.*, 2013). Expression data in DISCOVERY is from a study performed by DeRisi *et al.* in 2003; *The transcriptome of the intraerythrocytic developmental cycle of* ***Plasmodium falciparum*** (Mpangase *et al.*, 2013). The literature mining was done on PubMed's MEDLINE database abstracts (2012) by utilizing a form of the Aho-Corasick dictionary-matching algorithm. Articles are linked to proteins if they contain malaria-related keywords, and one of the protein's aliases. Additionally, abstracts are scanned for relational keywords, such as "interacts with", to positively or negatively connect proteins with chemical compounds from ChEMBL. Uncommon English words that occur frequently in articles linked to proteins are also stored (Mpangase *et al.*, 2013).

DISCOVERY2 also contains chemical compounds from the ChEMBL database with chemical search functionality and putative ligand-protein prediction information. These searches are performed using the MarvinSketch Applet, and JChemBase (ChemAxon) (Mpangase *et al.*, 2013). By using accession numbers, keywords or an advanced multi-parameter filtering interface can perform protein searches. An advanced search function was designed with the help of the Hibernate ORM (an object-relation mapper to provide a Java object interface to the relational database), together with a custom-built criteria library, to ensure the fastest possible filtering of protein sequences. Filtering can be done on

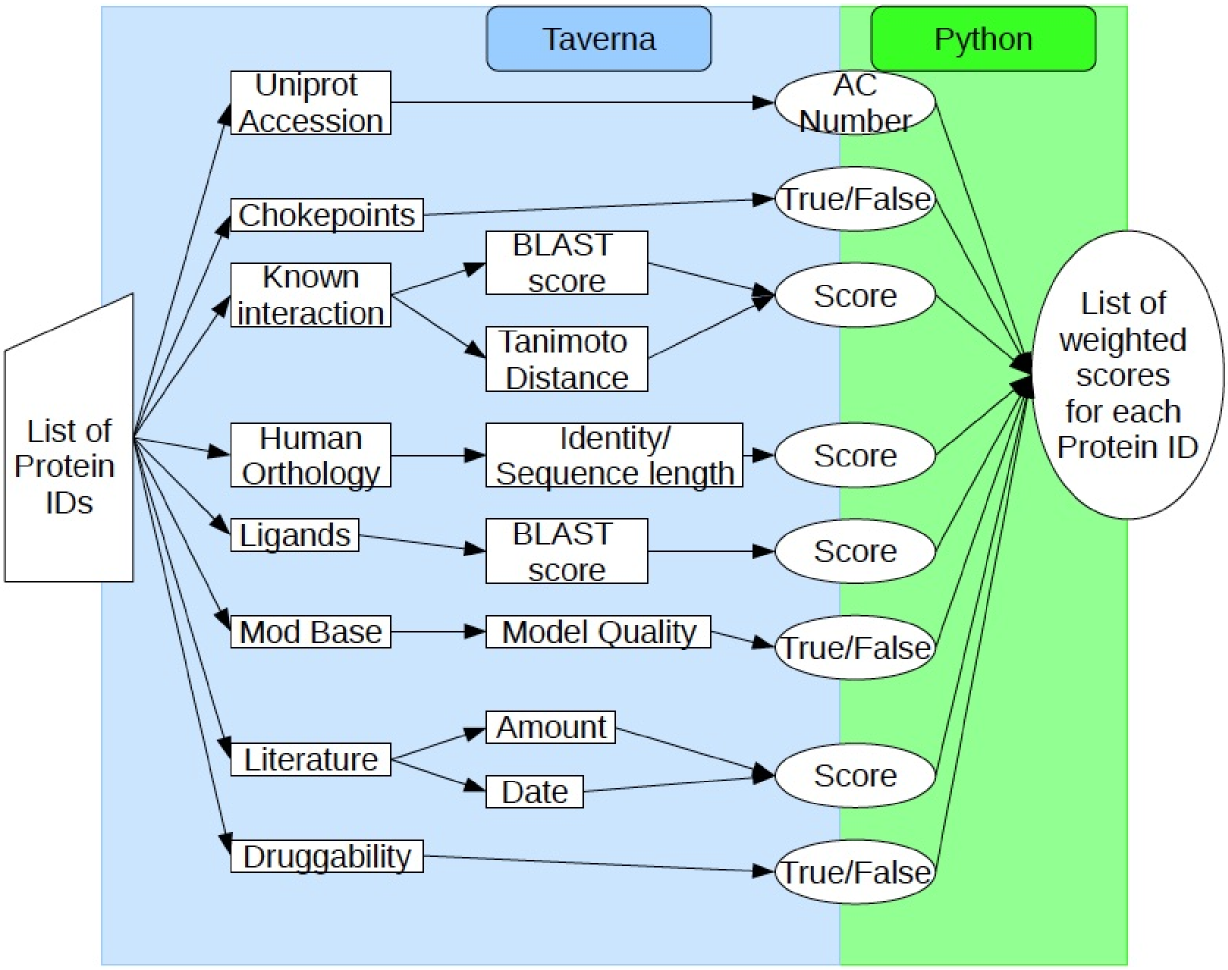


Figure 2.5: Flow diagram, from the *P. falciparum* input, through Taverna and Python, to the list of scores.

various fields including function, orthology and literature (Mpangase *et al.*, 2013).

2.2 Material and Methods

The aim to be achieved with the following methods is to identify proteins that have a possibility of being druggable. To achieve this the *P. falciparum* protein IDs were run through DISCOVERY, via web services using a Taverna pipeline.

Certain parts of the DISCOVERY database were wrapped as web services by J. Smit and used as functions in the Taverna Pipeline. All $\sim 5,500$ *P. falciparum* proteins were used as the input for the Taverna pipeline. Figure 2.5 illustrate the flow of the proteins, from a list of only protein IDs to a list of weighted scores.

2.2.1 Criteria considered

2.2.1.1 Uniprot Accession (AC) numbers

An accession number (AC) is assigned to each sequence upon inclusion into UniProtKB. Accession numbers are stable from release to release. If sequences are merged into one, for reasons of minimizing redundancy, the accession numbers of all relevant entries are kept, but each entry still has one primary AC and the rest are optional secondary ACs (UniProtFAQ, N/A). AC is stable identifiers and consists

Table 2.1: The format for AC numbers

1	2	3	4	5	6
[A-N, R-Z]	[0-9]	[A-Z]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]
[O, P, Q]	[0-9]	[A-Z, 0-9]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]

of six alphanumerical characters in the format as seen in Table 2.1. Examples are Examples are A2BC19, P12345.

The Primary Uniprot accession numbers are obtained for the proteins via DISCOVERY, as the AC numbers are stable from release to release. Even if the entries merge the AC numbers are kept. An AC is only deleted when the entry to which it was assigned has been removed from UniProtKB, but the list of deleted AC can be found in the document files (UniProtManual, 2014).

The UniProt accessions were assigned to the proteins in DISCOVERY using an identifier-mapping file downloaded from UniProt. The UniProt accessions were then used to assign GO annotations from UniProt-GOA database. Each GO term was linked to the Gene Ontology database and the AmiGO visualization tool is used to view the GO terms (Mpangase *et al.*, 2013).

The AC numbers are extracted from the DISCOVERY Database and used as a reference to be able to use it further down the line. These would not form part of the ranking and scoring of the proteins, but are simply part of the pipeline to have the data available to use in conjunction with the *P. falciparum* ID for future back referencing.

2.2.1.2 Chokepoints

Different enzymes in the metabolism can be good drug targets. These can be identified by chokepoints and the uniqueness of that pathway in the organism. "Chokepoint reactions" are defined by Yeh *et al.* (2004) "*...as a reaction that either uniquely consumes a specific substrate or uniquely produces a specific product in the PlasmoCyc metabolic network*" (Figure 2.6) (Yeh *et al.*, 2004). A chokepoint list was extracted from PlasmoCyc[1]. As proteins are only classified as a chokepoint or not a chokepoint (or unknown) this category only gave a true or false value as an output.

2.2.1.3 Known interactions

Glaxo Smith Kline (GSK) screened nearly 2 million compounds in its chemical library for inhibitors against malaria. Nearly $13,000$ compounds were confirmed to inhibit parasite growth by at least 80% at 2 microM concentration and more than $8,000$ compounds also showed potent activity against the multi-drug resistant strain Dd2. Approximately 82% of the compounds originated from the internal company projects, and are new to the malaria community. Analyses have suggested several novel mechanisms of antimalarial action, like the inhibition of protein kinases and host-pathogen interaction related targets. All these data were compiled into a database called TCAMS.

In DISCOVERY data for proteins-ligand interactions are obtained via the ChEMBL database. Doing a BLAST search against the ChEMBL protein targets generated these predictions and finding matching functional domains, these were generated through the InterProScan program. But only the data with 70% or more sequence similarity are shown in DISCOVERY.

To obtain the known interactions the *P. falciparum* proteins had to be compared with the ChEMBL proteins and scored by obtaining the percentage sequence similarity from the BLAST results. The

[1] http://plasmocyc.stanford.edu/uniqsubs.html, from which it was extracted, is no longer working, but the data can still be found at BioCyc.

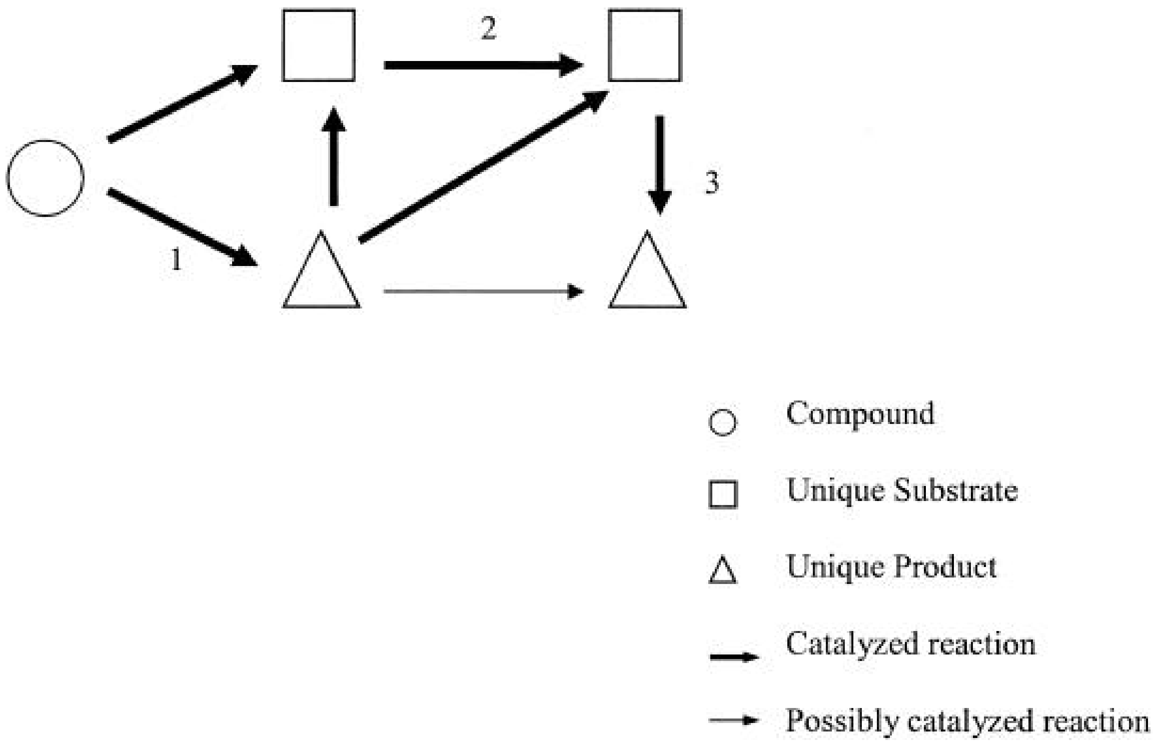


Figure 2.6: Diagram illustrating chokepoints.
The thick arrows represent reactions that are catalysed by enzymes, whereas the thin arrows represent reactions that are present, but with no evidence of the corresponding enzymes. When determining chokepoint reactions, only consider the catalysed reaction. (1) is a chokepoint, because it produces a unique product; (2) because it consumes a unique substrate and, (3) because it consumes a unique substrate and produces a unique product.
From: Yeh *et al.* (2004).

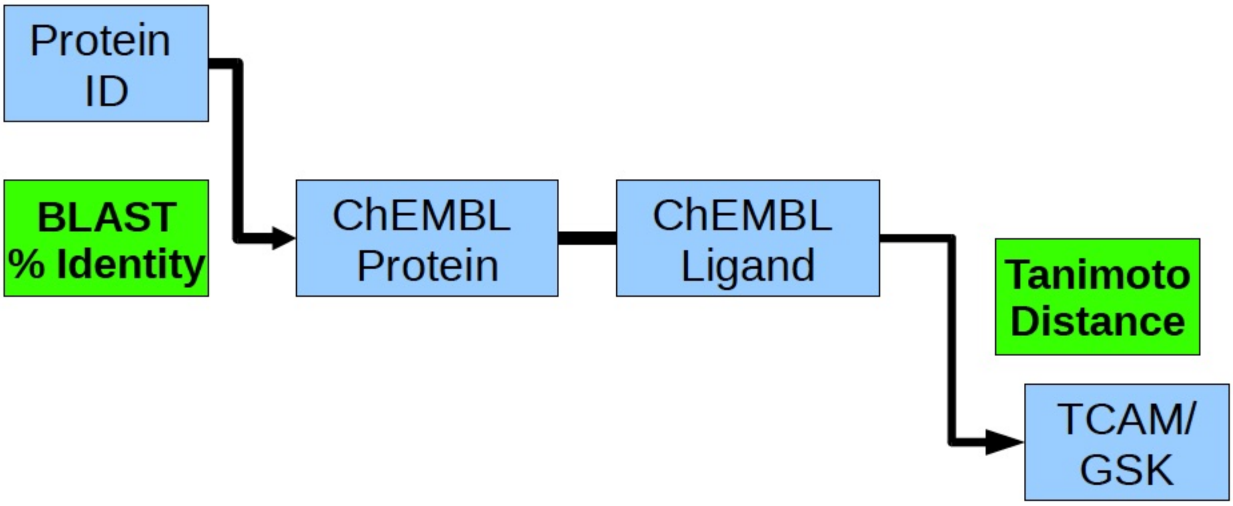


Figure 2.7: Diagram illustrating path of comparisons.

associated ChEMBL ligand had to be compared with the ligands from the TCAMS database and scored by calculating the Tanimoto distance (Figure 2.7). This caused a bit of difficulty as there is more than one ChEMBL protein similar to the *P. falciparum* protein and there are multiple ChEMBL ligands for each protein. For this it was decided to use the best scoring ChEMBL protein according to BLAST and its associated best scored ligand. Also to compare the ChEMBL ligands with the GSK ligands and calculate the Tanimoto distance wasn't as easy as expected. It was finally decided to compare all the GSK ligands with all the ChEMBL ligands in the DISCOVERY database and only store the ten best Tanimoto distances for each ChEMBL ligand.

The final scores and distances were calculated by working from the ligand to the protein: if there is a Tanimoto distance of more than 0.75 between a ChEMBL ligand and a GSK ligand then the output should also give the ChEMBL protein associated with that ligand and *P. falciparum* protein of that protein. Thus the output from this would be as follows (if the input would be the *P. falciparum* protein):

- ChEMBL protein;
- The E-value of the BLAST;
- The percentage sequence similarity between the two proteins according to the BLAST;
- ChEMBL ligand;
- GSK ligand and,
- Tanimoto distance.

From this output only the percentage similarity and the Tanimoto distance is of importance, the rest is only used as reference. To score this category the sequence similarity is "clustered" as follows:

- If it is more than 70%, it will be given 100%;
- If it is between 60% and 70%, it will be given 80%;
- If it is between 50% and 60%, it will be given 70%;
- If it is between 30% and 50%, it will be given 50% and,
- If it is less than 30%, the actual percentage will be used.

The Tanimoto distance was then used by multiplying it with this clustered score. If there is a Tanimoto distance, and it is less than 0.75, for a specific *P. falciparum* the output for this pipe will be zero.

2.2.1.4 Human orthologs

A human ortholog needs to be identified because if there is no ortholog, or a very dissimilar ortholog, between the parasite and humans it is less likely that the drug will have a toxic effect on the human.

Orthologs are identified in DISCOVERY via a T-Coffee alignment. During the pipeline all human orthologs are given as an output together with a score to help with the prediction of similarity. From this only the ortholog with the smallest "similarity score" is used. As % change is not a good indicator of similarity it was decided to rather use the following: Identity divided by Sequence length (according to the T-coffee output). If there are no human orthologs for a specific *P. falciparum* protein then the protein will get full marks for the "Human orthologs" score, otherwise the score will increase logarithmically, with the decrease in identity/sequence length. The score will always be negative due to the value of Identity over Sequence length always being between 0 and 1. Thus the absolute log value is taken as the score, giving an equation as can be seen in Equation 2.1:

$$Score = \mid log(\frac{Identity}{SequenceLength}) \mid \tag{2.1}$$

When Identity over sequence lenght tend to be toward 0, then the log value will tend toward $-\infty$, thus the absolute value will tend toward ∞. In these cases the protein will get full marks.

2.2.1.5 Ligands

Orthologous proteins might interact with known ligands or there might be similar ligands to known antimalarial drugs, all these might make suitable drugs. This is similar to “Known interactions” but without the Tanimoto distance part (see Figure 2.7). In other words, the *P. falciparum* protein is compared to ChEMBL proteins that have associated ligands. Added to this is the knowledge of the *P. falciparum* protein having domains also found in other proteins. This means that part of this pipe is similar, almost identical, to “Known interactions”. The comparison of *P. falciparum* to ChEMBL proteins is done using BLAST and comprised 70% of this score; it was clustered and scored as follows:

- If it is more than 70%, it was given 100%, of which 70% is 70;
- If it is between 60% and 70%, it was given 80%, of which 70% is 56;
- If it is between 50% and 60%, it was given 70%, of which 70% is 49;
- If it is between 30% and 50%, it was given 50%, of which 70% is 35 and,
- If it is less than 30%, the actual percentage was used, which caused the score to range from zero to 21.

Knowledge of the *P. falciparum* domains will account for 30% of the score. As the output for the domains are true or false it will be given a score of 30 if it is true, and a score of zero if it is false. These two scores will then be added to get a total potential score up to 100.

2.2.1.6 ModBase structure

Protein structures can be beneficial if the protein is used in further research. ModBase is a database of annotated protein structure models, which is derived from ModPipe. . Amongst other scores, ModBase gives a Discrete Optimized Protein Energy (DOPE) score. It is said that if the DOPE score is less than zero the model is reliable and more than zero is an unreliable model. Thus, in the pipeline the smallest DOPE score is given if there are a few potential models. These are scored in an all or nothing manner, meaning if the DOPE score is negative 100% will be given, it doesn’t matter how negative it is. If it is positive 0% will be given.

2.2.1.7 Literature

Protein ligand interactions might be depicted in literature. These interactions can either be positive (the ligand inhibits the protein) or they might be negative (the ligand enhances the protein). Literature can be mined to identify these interactions, but the risk of false positives or false negatives can be high. In this part of the pipeline it was only relevant if there has been any research done on this specific protein and how old the research is, being beneficial for future research if if there are articles available. Older research are not considered less important, but if a protein was not used in recent research and with techniques advancing rapidly, the question of why no new research are being done might need to be asked. Thus the scoring for this will be as follows:

- More than three articles, and the latest one less than five years old will be given a score of 100%;
- Less than three articles, and the latest one less than five years old will be given a score of 75%;
- More than three articles, and the latest one is between five and ten years old will be given a score of 75%;
- Less than three articles, and the latest one is between five and ten years old will be given a score of 50%;
- Articles published, but they are all older than ten years will be given a score of 25% and,
- If there are no published articles no score will be awarded.

These scores are according to the latest PubMed abstracts.

2.2.1.8 Druggability

Druggability is a term used to describe a biological target, such as a protein, that is known to, or is predicted to, bind with high affinity to a drug. The binding of the drug to a druggable target must alter the function of the target with a therapeutic benefit to the patient. Druggability in DISCOVERY is calculated by considering DrugEBIlity. DrugEBIlity is a structure-based druggability search engine at ChEMBL. In DrugEBIlity users can survey different types of druggability scores of a given protein structure. This can be done as follows:

- If one have a px number (SCOP domain) or a pdb code, its associated data can be accessed by clicking on the “3D Domain” link on the DrugEBIlity site;
- If one have an accession number, more information can be accessed by clicking on the “Protein” link on the DrugEBIlity site;
- If one have a sequence, it can be run against all domain or gene sequences by using the “BLAST” link on the DrugEBIlity site and,
- By uploading a structure on the DrugEBIlity site users can calculate structure-based druggability scores.

Druggability can be predicted by using different methods that may rely on evolutionary relationships, 3D-structural properties, or other descriptors. Druggability in the pipeline is identified as a true or false value; it is druggable or it is not. This information is gathered from the DrugEBIlity site by comparing *Plasmodium falciparum* proteins to known druggable proteins. Consequently this criterion is again scored on an all-or-nothing basis.

2.2.2 Scoring and Weighting

Little known information was collected regarding the scoring of potential drug targets. Some databases that works on a scoring system, like TDR, uses the users own discretion thus the user can choose its own weights in scoring the proteins. For this reason an in-house scoring system has been developed to score the proteins based on what might be an important factor in a protein. The following criteria were considered: known structure, druggability, essential for survival (chokepoint), human homologs, activity in diseased stated etc. Known interactions, ligands that bind to the protein and literature information was added as well while activity in diseased state was removed due to a lack of ability to process the information. Some important criteria have been scored lower due to little known information and uncertainty regarding the data’s correctness.

Proteins are not scored negatively, due to the data not always being available. If the criteria are undesirable or the data is unknown the protein will simply not get a score for that criteria. If it is known and desirable the protein will get a full score or a partial score for that criteria, depending on the criteria and how desirable the data of that protein is. All scores are calculated out of 10 and then weighted according to its "priority". Top priority is given to chokepoints due to it being considered an important characteristic for drug targets. Druggability is scored at the bottom due to unknown data. Literature and ModBase is not considered as important as it only eases future research to be done on the protein. The scores for all seven criteria are as follows:

- Chokepoints are given top priority at 70, with an "all or nothing" score. If it is a chokepoint, according to Stanford, it is given full scores, otherwise no score;
- Known interactions are given a weighted score between 0 and 60, based on its BLAST percentage sequence similarity multiplied by its Tanimoto distance. These scores are then clustered;
- Human orthologs are also given weighted score between 0 and 50, based on the absolute logarithm of the identity divided by the sequence length;
- Ligands are again given a weighted score between 0 and 40, based on the BLAST percentage sequence similarity clustered and added to the "all or nothing" score of the domain matches;
- ModBase is given an "all or nothing" score with a final value of 30;
- Literature is given clustered values at a final maximum value of 20 and,
- Druggability, due to a lot of unknown data, is scored the lowest at an "all or nothing" score of 10.

All weighted scores are then added for a protein and divided by 280 (the maximum potential value) to get a final aggregate score between 0 and 1.

2.3 Results

2.3.1 Calculation of statistical parameters

2.3.1.1 Chokepoints

Chokepoint reactions can be defined as "*...as a reaction that either uniquely consumes a specific substrate or uniquely produces a specific product in the PlasmoCyc metabolic network*" (Yeh *et al.*, 2004). A protein is either considered a chokepoint or not. And is seen as the most important criteria as many proteins that are considered as drug targets are chokepoints. Thus it will have a total weighted score of 70.

2.3.1.2 Known interactions

By comparing malaria proteins to proteins with known ligands, and in turn compare these ligands to ligands from the GSK database, that is known to inhibit the malaria parasite; a high potential antimalarial drug target can be identified. Thus this score will have the second highest weight at 60.

2.3.1.3 Human orthologs

If malarial proteins don't have human orthologs or these orthologs are very dissimilar a potential drug target can be identified by reducing toxicity against the drug in the human host. Due to some proteins making successful drug targets despite the potential toxicity in humans this criteria is still considered important, but not as important as Chokepoints and Known interactions, thus it will have a weight of 50.

2.3.1.4 Ligands

Ligands is obtained by comparing malaria proteins with proteins with known ligands. This can aid in identifiying potential drug targets. This score will have a weight of 40.

2.3.1.5 ModBase

ModBase is a database of annotated protein structure modules. By using a DOPE score proteins can be identified as having a structure (DOPE score is less than zero) or not having structure (DOPE score is more than zero). Because protein structures can be beneficial for future research this criteria is not weighted heavily and caries a weight of 30.

2.3.1.6 Literature

Knowledge and the amount of knowledge available for a specific protein can again aid in future research and because of this Literature only carries a weight of 20.

2.3.1.7 Druggability

Druggability can be predicted by using different methods that may rely on evolutionary relationships, 3D-structural properties, or other descriptors. In this study a protein is considered as druggable or not. Due to uncertainity Druggability is not weighted heavily, it is only carries a weight of 10. As more information becomes available this weight can be revised.

2.3.1.8 Final aggregate score

The final score is calculated by adding all the scores for the different criteria and dividing it by 280, which is the maximum potential score for a protein. This will result in a score between zero and one.

2.3.2 Creation and evaluation of workflows

Due to the vast amount of data the protein IDs had to be split into two bins. Due to the amount of time it took for the “known interactions” to run through Taverna and causing time-outs and errors in the Taverna runs, the “Known interaction” part was run directly on the server. Thus, only “Uniprot accession numbers”, “Chokepoints”, “Orthologs”, “Ligands”, “ModBase”, “Literature” and “Druggability” were received as output (in two bins) when running the two protein ID bins through Taverna. These were then saved as excel sheets and the results look like Table 2.2.

2.3.3 Calculation of “Final aggregate score”

A Python program was written that would combine all the pieces of data together (a list of protein ID’s, the “Known interaction” output and the two bins containing the rest of the Taverna outputs). The Python program also scored the data and output a csv file containing only the scores and the protein IDs and a csv file that contains all the most important data. These csv files were then sorted according to the “Final aggregate score”. A portion of the csv file containing only the scores can be seen in Table 2.3 and a portion of the entire table can be seen in Table 2.4.

The top 10 ranked scores can be seen in Table 2.5. They are:

1. PFC0855w with an aggregate score of 0.96;
2. PFD0830w with an aggregate score of 0.92;

Table 2.2: Taverna results from one of the two runs.

The Taverna input (protein IDs) were split into two, thus the Taverna output is also spit into two. Here are the first ten results of the first "bin".

PID	MAL13	MAL13	MAL13	MAL13	MAL13	MAL13	MAL13	MAL13	MAL13	MAL13
Values	P1.1	P1.100	P1.102	P1.103	P1.105	P1.106	P1.107	P1.11	P1.111	P1.112
Druggability	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Articles	167	0	0	0	0	0	0	0	0	0
	2011									
	2011									
	2011									
	2011									
	2010									
Chokepoint	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ModBase	1.23	−0.1			−2.1		−0.03		−0.92	
Orthologs					ENSP00000311344|0.14452215					
					ENSP00000324804|0.13752913					
					ENSP00000343317|0.12470862					
					ENSP00000375668|0.102564104					
					ENSP00000376775|0.11655012					
					ENSP00000311344|0.14452215					
					ENSP00000396821|0.13519813					
					ENSP00000403301|0.11072261					
					ENSP00000410671|0.12237762					
					ENSP00000415067|0.11305361					
					ENSP00000415759|0.1013986					
					ENSP00000437193|0.14219114					
Ligands	0 | false	0 | true	0 | false	0 | false	32% | true	0 | false	0 | false	0 | false	0 | false	0 | false
UniProt	Q8IEV1		Q8IEC9	Q8IEC2	Q8IEB9	Q8IEB8	Q8IEB7		Q8IEB2	

Table 2.3: Ten entries from the score.csv file.

Via Python all the data pieces were incorporated into one file, but containing only the scores for each point and the final aggregate score.

PID	MAL13 P1.1	MAL13 P1.100	MAL13 P1.102	MAL13 P1.103	MAL13 P1.105	MAL13 P1.106	MAL13 P1.107	MAL13 P1.11	MAL13 P1.111	MAL13 P1.112
Chokepoint score	0	0	0	0	0	0	0	0	0	0
Interaction score	0	0	0	0	3.84	0	0	0	0	0
Ortholog score	10	10	10	10	8.4	10	10	10	10	10
Ligand score	0	3	0	0	6.5	0	0	0	0	0
ModBase score	0	10	0	0	10	0	10	0	10	0
Literature score	10	0	0	0	0	0	0	0	0	0
Druggability score	0	0	0	0	0	0	0	0	0	0
Total score	70	92	50	50	121.1	50	80	50	80	50
Weighted score	0.25	0.33	0.18	0.18	0.43	0.18	0.29	0.18	0.29	0.18

Table 2.4: Ten entries from alldata.csv file.

Via Python all the data pieces were incorporated into one file, containing the scores for each point, the final aggregate score, and the pieces of information that the scores depend on.

PID	MAL13 P1.1	MAL13 P1.100	MAL13 P1.102	MAL13 P1.103	MAL13 P1.105	MAL13 P1.106	MAL13 P1.107	MAL13 P1.11	MAL13 P1.111	MAL13 P1.112
UniProt	Q8IEV1		Q8IEC9	Q8IEC2	Q8IEB9	Q8IEB8	Q8IEB7		Q8IEB2	
Chokepoint	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Chokepoint score	0	0	0	0	0	0	0	0	0	0
ChEMBL Protein					ChEMBL 3557					
ChEMBL Ligand					ChEMBL 8119					
GSK Ligand					ChEMBL 630301					
Identity	0	0	0	0	32	0	0	0	0	0
Tanimoto Distance	0	0	0	0	0.77	0	0	0	0	0
Interaction Score	0	0	0	0	3.84	0	0	0	0	0
Ortholog ID					ENSP00000311344					
Highest value	0	0	0	0	0.14	0	0	0	0	0
Ortholog score	10	10	10	10	8.4	10	10	10	10	10
Similarity	0	0	0	0	32	0	0	0	0	0
Domain matches	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
Ligand score	0	3	0	0	6.5	0	0	0	0	0
ModBase	1.23	-0.1			-2.1		-0.03		-0.92	
ModBase score	0	10	0	0	10	0	10	0	10	0
Articles	167	0	0	0	0	0	0	0	0	0
Date	2011									
Literature score	10	0	0	0	0	0	0	0	0	0
Druggability	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Druggability score	0	0	0	0	0	0	0	0	0	0
Final aggregate score	70	92	50	50	121.1	50	80	50	80	50
Weighted score	0.25	0.33	0.18	0.18	0.43	0.18	0.29	0.18	0.29	0.18

Table 2.5: Top ten ranked PID with the corresponding scores. The PIDs are sorted via the aggregate score, giving the top ten ranking proteins.

PID	PFC 0855w	PFD 0830w	PF08_0066	PF08_0034	PFE 0660c	PF13_0128	PFB 0685c	PFA 0145c	PFI 1005w	PF10_0340
Chokepoint score	10	10	10	10	10	10	10	10	10	10
Interaction score	10	10	5.7	8.8	5.2	8.3	6.4	10	4.3	10
Ortholog score	10	5.72	10	10	10	10	10	4.42	10	7.11
Ligand score	10	10	10	10	10	10	8.6	10	10	10
ModBase score	10	10	10	10	10	10	10	10	10	10
Literature score	5	10	10	0	10	0	7.5	7.5	10	0
Druggability score	10	10	10	10	10	10	10	10	10	10
Total score	270	258.6	254.2	252.8	251.2	249.8	247.8	247.1	245.8	245.55
Weighted score	0.96	0.92	0.91	0.90	0.90	0.89	0.89	0.88	0.88	0.88

3. PF08_0066 with an aggregate score of 0.91;
4. PF08_0034 with an aggregate score of 0.9;
5. PFE0660c with an aggregate score of 0.9;
6. PF13_0128 with an aggregate score of 0.89;
7. PFB0685c with an aggregate score of 0.89;
8. PFA0145c with an aggregate score of 0.88;
9. PFI1005w with an aggregate score of 0.88 and,
10. PF10_0340 with an aggregate score of 0.88.

2.3.4 Statistical evaluation

In R the statistical distribution of the data was calculated for the different scores as well as for the "Final aggregate score". To see if a specific trend is followed distribution histograms were made available via R.

2.3.4.1 Chokepoints

For "Chokepoint" the following values were identified using R:

- Minimum value is zero (0);
- The first quarter value is also zero (0);
- The median has a value of zero (0);
- The mean has a value of 0.4626;
- The third quarter at a value of zero (0) and,
- The maximum value is ten (10).

A histogram of the "Chokepoint" data distribution can be seen in Figure 2.8A. Due to "Chokepoint" being an all or nothing criteria it is expected to have a minimum value of zero and a maximum value of ten. It is also expected that the first quarter, median and third quarter will either be zero or ten. The data are skewed towards zero, which can be due to predicting most of the proteins not as chokepoints. This can be correct, or erroneous due to lack of information.

2.3.4.2 Known interactions

For "Known interactions" the following values were identified using R:

- Minimum value is zero (0);
- The first quarter value is also zero (0);
- The median has a value of zero (0);
- The mean has a value of 0.6053;
- The third quarter at a value of zero (0) and,
- The maximum value is ten (10).

A histogram of the "Known interaction" data distribution can be seen in Figure 2.8B. Even though "Known interactions" can take on any value between zero and ten, it is still skewed toward zero, with less than 700 of the approximate 5, 500 proteins having scores more than zero. Thus, the first quarter, median and third quarter has values of zeros. This skew toward zero can again be due to a lack of information, but is also likely due to the proteins not having any known ligands that can bind to it.

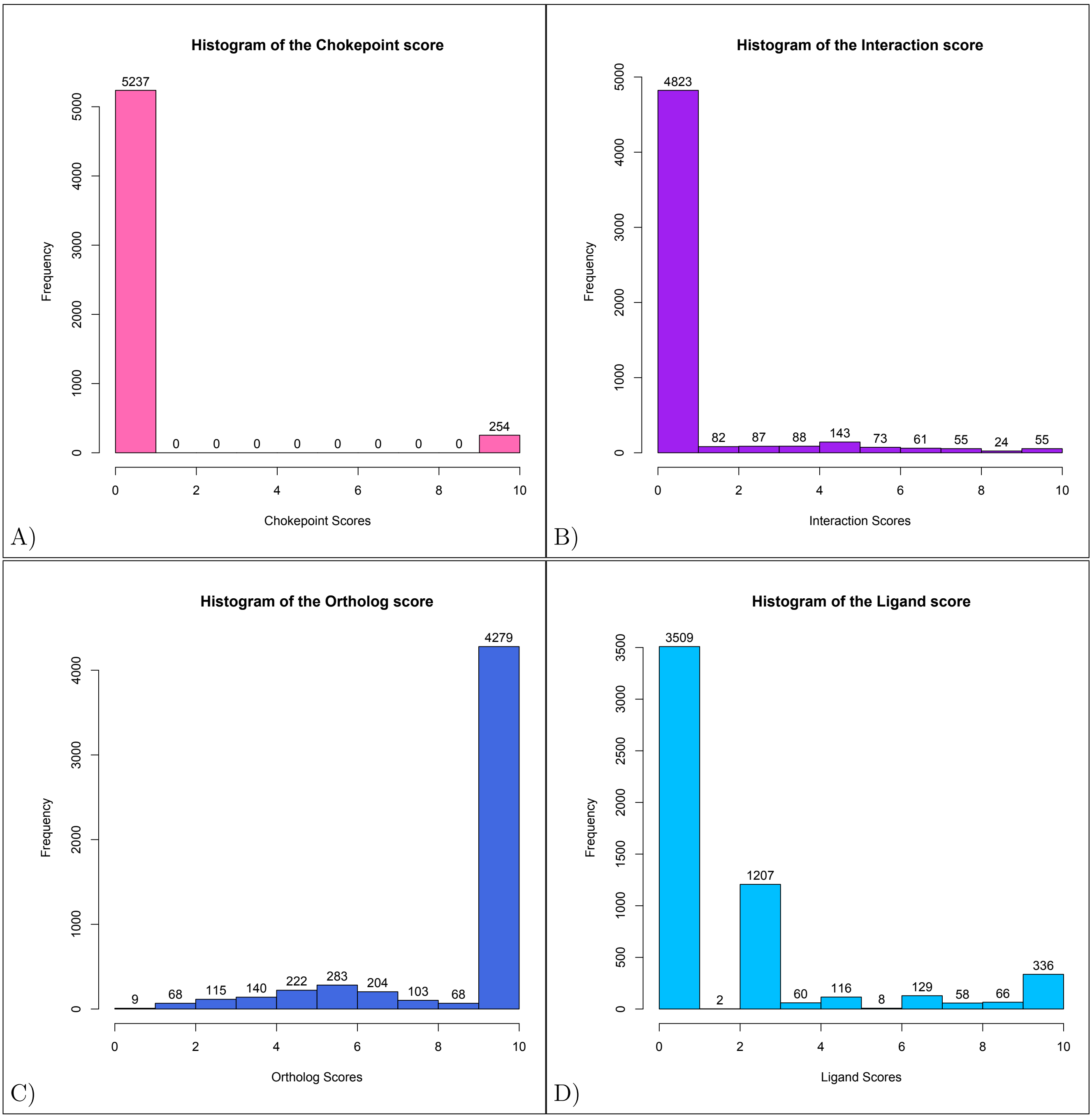


Figure 2.8: Histogram of the distribution of some of the scores.

(A) Histogram showing the distribution of the "Chokepoint" scores.

(B) Histogram showing the distribution of the "Known interaction" scores.

(C) Histogram showing the distribution of the "Orthologs" scores.

(D) Histogram showing the distribution of the "Ligand" scores.

2.3.4.3 Orthologs

For the “Orthologs” the following values were identified using R:

- Minimum value is 0.07;
- The first quarter value is also ten (10);
- The median has a value of ten (10);
- The mean has a value of 8.902;
- The third quarter at a value of ten (10) and,
- The maximum value is ten (10).

A histogram of the “Orthologs” data distribution can be seen in Figure 2.8C. “Orthologs” can also take on any value between zero and ten, but here it is skewed toward ten with 1,240 of the approximate 5,500 proteins not having a full score. This will result in the first quarter, median and last quarter having a value of ten. This skew toward ten can be due to inadequate information or to a genuine lack of orthologs between humans and the *P. falciparum* proteins.

2.3.4.4 Ligands

For “Ligands” the following values were identified using R:

- Minimum value is zero (0);
- The first quarter value is also zero (0);
- The median has a value of zero (0);
- The mean has a value of 1.754;
- The third quarter at a value of three (3) and,
- The maximum value is ten (10).

A histogram of the “Ligands” data distribution can be seen in Figure 2.8D. “Ligands” can again take on any value between zero and ten. It is again slightly skewed toward zero, but less than in the previous two cases, resulting in a first quarter and median of zero, but the third quarter had a value other than the minimum value (zero) or maximum value (ten). Again, this skew can be due to inadequate information, but also, actually, due to no related ChEMBL proteins when comparing it via a BLAST search.

2.3.4.5 ModBase

For the “ModBase” scores the following values were identified using R:

- Minimum value is zero (0);
- The first quarter value is also zero (0);
- The median has a value of zero (0);
- The mean has a value of 4.103;
- The third quarter at a value of ten (10) and,
- The maximum value is ten (10).

A histogram of the "ModBase" data distribution can be seen in Figure 2.9A. "ModBase" has, again, an all or nothing score, meaning it will be either zero or ten. Here the "ModBase" scores are roughly equally divided between zero (3, 238 proteins scored zero) and ten (2, 253 proteins scored 10), but slightly skewed toward zero, resulting in the first quarter and median having a value of zero, but the third quarter's value is ten. This skew is likely due to inadequate information, as it is not an easy task to be able to get a 3D structure for a protein.

2.3.4.6 Literature

For the "Literature" scores the following values were identified using R:

- Minimum value is zero (0);
- The first quarter value is also zero (0);
- The median has a value of zero (0);
- The mean has a value of 1.261;
- The third quarter at a value of zero (0) and,
- The maximum value is ten (10).

A histogram of the "Literature" data distribution can be seen in Figure 2.9B. "Literature" do not have a continuous or an all-or-nothing score, as the previous criteria, "Literature" can either score zero, 2.5, five, 7.5 or ten; depending on the amount of articles available for that protein and the date of the last published article. Again it is skewed toward zero, with 4, 621 proteins having no articles, resulting in a first quarter, median and third quarter with a value of zero. This skew is most likely due to research that has not yet been done on a specific protein.

2.3.4.7 Druggability

For "Druggability" the following values were identified using R:

- Minimum value is zero (0);
- The first quarter value is also zero (0);
- The median has a value of zero (0);
- The mean has a value of 1.783;
- The third quarter at a value of zero (0) and,
- The maximum value is ten (10).

A histogram of the "Druggability" data distribution can be seen in Figure 2.9C. Finally, "Druggability" has again an all or nothing score. Thus, all the values should be zero (4, 512 proteins) or ten (979 proteins), and in this case; as it is skewed toward zero again; the first quarter, median and third quarter have a value of zero. This skew is either due to inadequate information or due to the protein not having the properties or structure to be likely to bind to small molecules.

2.3.4.8 Final aggregate score

For the "Final aggregate" scores the following values were identified using R:

- Minimum value is 0.01;
- The first quarter value is also 0.18;
- The median has a value of 0.22;

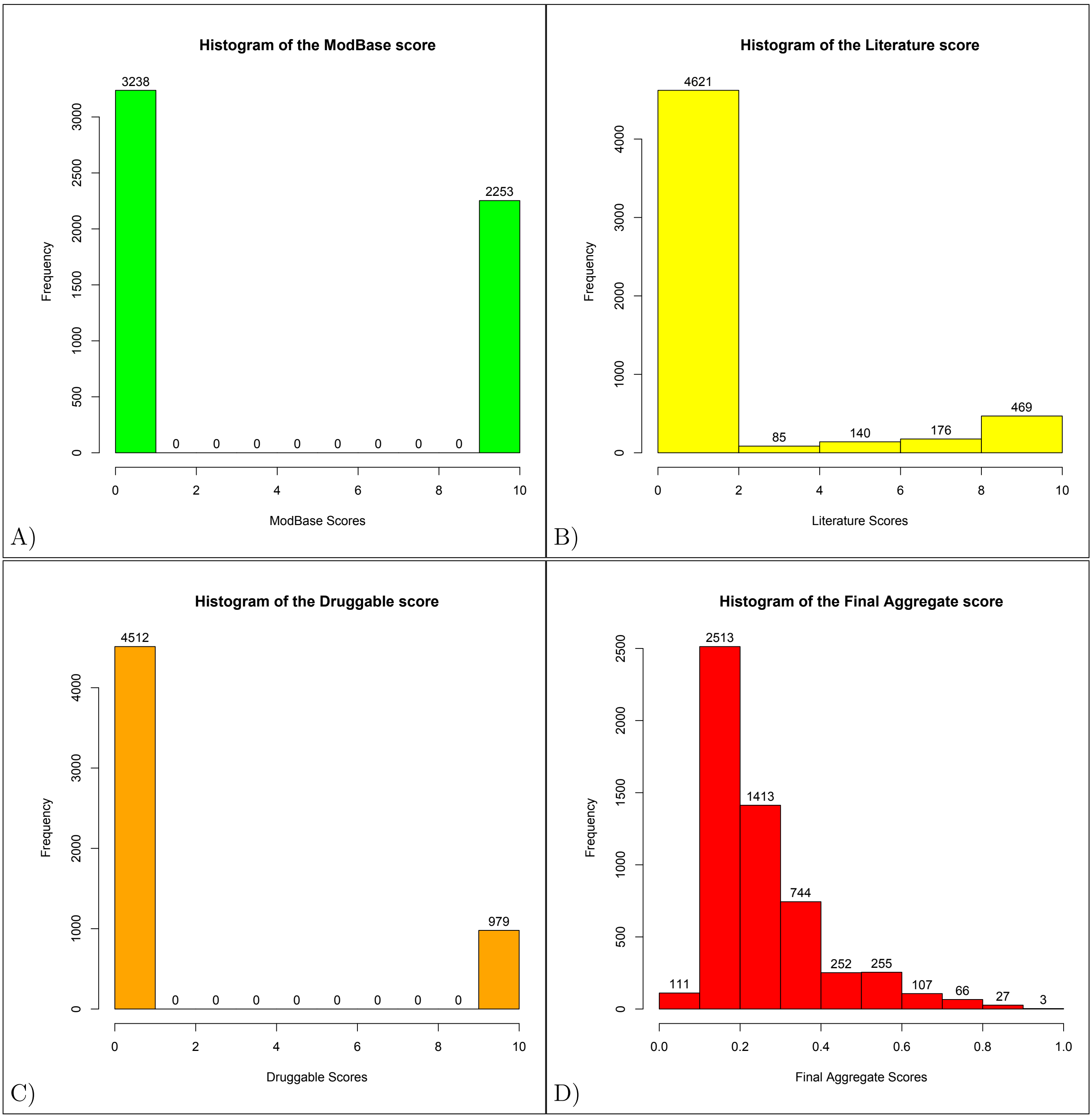


Figure 2.9: Histogram of the distribution of the rest of the scores.

(A) Histogram showing the distribution of the “ModBase” scores.

(B) Histogram showing the distribution of the “Literature” scores.

(C) Histogram showing the distribution of the “Druggability” scores.

(D) Histogram showing the distribution of the “Final aggregate” scores.

Table 2.6: Percentiles for the "Final aggregate score" as identified by R.

Percentile	1%	5%	10%	25%	50%	75%	90%	95%	99%
Score	0.08	0.17	0.18	0.18	0.22	0.32	0.47	0.57	0.77

- The mean has a value of 0.2688;
- The third quarter at a value of 0.32 and,
- The maximum value is 0.96.

A histogram of the "Final aggregate" score distribution can be seen in Figure 2.9D. Lastly, the "Final aggregate" score can have any score between zero and one, but is in effect between 0.01 and 0.96. All these "skewed toward zero" of the different criteria had made the final score skewed positively with a long tail towards the right. It peaks around 0.18, which is also the value of the first quarter. The value of the median is 0.22 and that of the third quarter is 0.32. Due to the fact that the "Final aggregate score" histogram do not have a normal distribution calculating a significant score was tricky. To calculate a significant score percentiles was used and R identified values as can be seen in Table 2.6.

Using Table 2.6, scores equal or bigger than 0.77 can be considered a good score as only 1% of all the aggregate scores are equal or bigger than this value, or 99% of all values are smaller. This came to 58 possible proteins.

2.4 Discussion

All 5,491 proteins of the *P. falciparum* were run through DISCOVERY via the Taverna pipeline these outputs were then scored via a Python program. The scores were ordered in a descending order. Some of the *Plasmodium* proteins showed promising scores as these had scores bigger than 0.77. These included three proteins that have been identified as a drug target for malaria a while back, some dating from early 1970s (Triglia *et al.*, 1997). Namely:

- 1-deoxy-D-xylulose-5-phosphate reductoisomerase, identified by PF14_0641. This is a target for fosmidomycin, and it scored 0.84 (Jomaa *et al.*, 1999);
- Dihydrofolate reductase it is identified by PFD0830w. It is a target for pyrimethamine and cycloguanil and had a score of 0.92 (Yuthavong *et al.*, 2012). And,
- Dihydropteroate synthase, identified by PF08_0095. It is a target for sulfone/sulfonamide drugs and had a score of 0.79 (Triglia *et al.*, 1997).

PFC0855w, a putative protein that is an ubiquitin-conjugating enzyme, is currently identified on the PlasmoDB site by PF3D7_0319300 (it was previously also identified by MAL3P6.32). This protein had the highest score at 0.96, with a perfect 10 for all the criteria except literature, where it scores a 5, meaning that:

- It is a chokepoint according to Stanford;
- It has a sequence similarity of 76% with the ChEMBL protein ChEMBL6089;
    - Also ChEMBL6089 has a ligand, ChEMBL86464 that has a Tanimoto distance of one with the GSK ligand, ChEMBL239794;
- It does not have a known human ortholog;

- It has a sequence similarity of 76% to a ChEMBL protein and it also has known domains;
- It has a ModBase score of −2.1;
- There are 2 articles published, mentioning this protein, of which the latest one was published in 2007 and,
- It is also considered druggable.

PF10_0179a identifies the protein that had the lowest score of 0.1. This protein only had a score for "orthologs". Thus:

- It is not a chokepoint;
- It has no similarity to any known other sequences and thus also have no ligand that have a possibility of binding to this protein;
- It do have a human ortholog, and this have a score of 0.72 out of the possible 10 points, thus it was considered quite similar;
- It still has no sequence similarity;
- There are no ModBase score;
- There are no articles ever published on this protein and,
- This protein is not considered druggable.

The results for the highest and lowest protein look quite different. The reason PF10_0179a scored so low might be due to the fact that there is not a lot of research yet done on this protein. And there is quite a bit done on PFC0855w, even though the article score for this protein is moderate.

2.5 Conclusion

Even though a lot of the the data from the individual criteria are skewed toward zero, and even the "Final aggregate" score is skewed toward zero, the results obtained from this in-house scoring technique shows that there are a few proteins that show promise as the next drug target for *P. falciparum.* 58 out of the possible $5,491$ proteins (1%) had a score of 0.77 or bigger. This 1% of proteins can be considered to have a good score. The three known drug targets are included in these 58 proteins with respective scores of 0.79, 0.84 and 0.92, reassuring that the choice of the scoring and weighting techniques can be plausible to implement in the identification of future drug targets for *P. falciparum.*

Chapter 3

Case Studies

3.1 Introduction

Case studies can be defined as "*a process or record of research in which detailed consideration is given to the development of a particular person, group, or situation over a period of time*". In this instance consideration is given to four *P. falciparum* proteins as observed in the DISCOVERY database, as well as after the protein was scored using the Taverna pipeline and Python program.

Case studies were chosen to illustrate the comparison of a protein between the results observed in DISCOVERY and the scores observed from the Taverna output, and also to illustrate the results observed between proteins. Thus four proteins were chosen with remarkably different aggregate scores; a very low score, as close to zero as possible, but it still needs to have some information available; a very high score, as close to one as possible; and two scores as evenly distributed, between the other two scores, as possible. Proteins chosen for case studies should preferably not be one of the three proteins known to be antimalarial drug targets. The four proteins were:

1. PF11_0058-b: RNA polymerase subunit, putative. With a score of 0.16 it serves as the very low score;
2. PFL2280w: Serine/Threonine protein kinase, putative. With a score of 0.5 it is one of proteins with an intermediate score;
3. PFE1050w: S-adenosyl-L-homocysteine hydrolase. It has a score of 0.77 and thus is the second intermediate protein and,
4. PFC0855w: Ubiquitin conjugating enzyme, putative. This protein had the highest score observed after Taverna and Python, thus at a score of 0.96 it was chosen as the very high score protein to discuss in the case study.

3.2 PF11_0058-b: RNA polymerase subunit, putative

3.2.1 Introduction

Transcription in eukaryotes and prokaryotes functions by the same fundamental mechanisms in all cells, but it is considerably more complex in eukaryotic cells compared to bacteria. This increased complexity of eukaryotic transcription is likely to facilitate the sophisticated regulation of gene expression needed to direct the activities of the many different cell types of multicellular organisms (Cooper, 2000).

All three of the nuclear RNA Polymerases (RPs) are complex enzymes and consists of 8 to 14 different subunits each. Although they recognize different promoters and transcribe distinct classes of

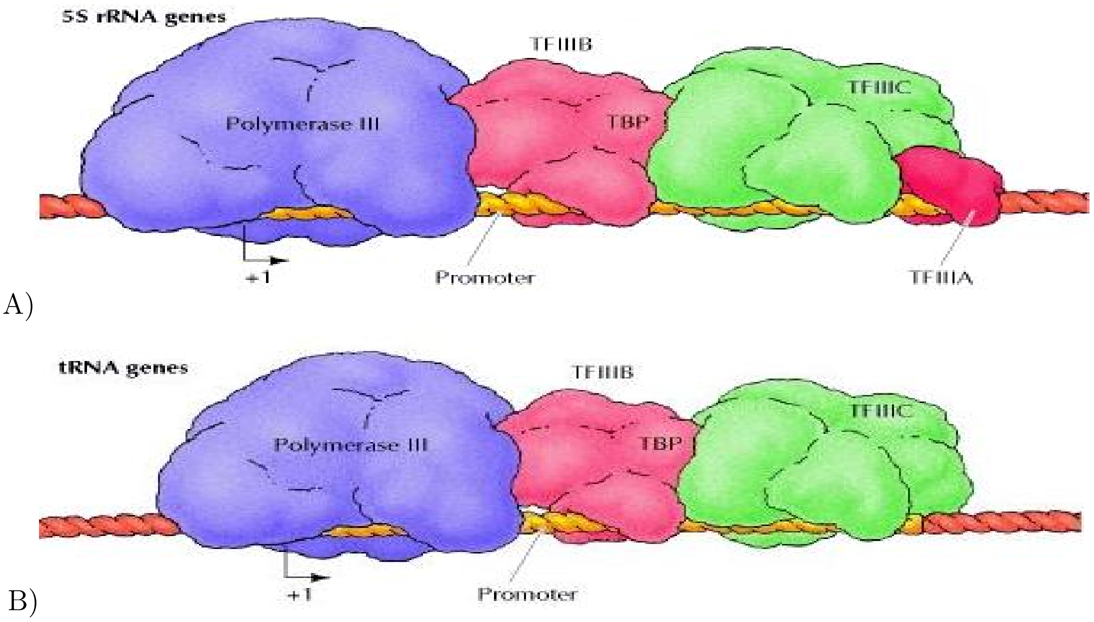


Figure 3.1: Transcription of polymerase III genes.
The promoters of 5S rRNA and tRNA genes are downstream of the transcription initiation site.
A) Transcription of the 5S rRNA gene is initiated by the binding of TFIIIA, followed by the binding of TFIIIC, TFIIIB, and the polymerase.
B) The tRNA promoters do not contain a binding site for TFIIIA, and TFIIIA is not required for their transcription.
Instead, TFIIIC initiates the transcription of tRNA genes by binding to promoter sequences, followed by the association of TFIIIB and polymerase. The TBP is a subunit of TFIIIB.
From: Cooper (2000).

genes, they share several common features. The two largest subunits of all three eukaryotic RPs are related to the β and β' subunits of the single *E. coli* RP. In addition, five subunits of the eukaryotic RPs are common to all three different enzymes. Consistent with these structural similarities, the different eukaryotic RPs share several functional properties, including the need to interact with other proteins to appropriately initiate transcription (Cooper, 2000).

Distinct RPs are responsible for the transcription of genes encoding rRNA and tRNA in eukaryotic cells. All three RPs, however, require additional transcription factors (TF) to associate with appropriate promoter sequences. The three different polymerases in eukaryotic cells recognize distinct types of promoters but a common TF, namely the TATA-binding protein (TBP), appears to be required for the initiation of transcription by all three enzymes (Cooper, 2000).

3.2.1.1 RNA polymerase I

RP I is devoted solely to the transcription of rRNA genes, which are present in tandem repeats. Transcription of these genes yields a large 45S pre-rRNA, which is then processed to yield the 28S, 18S, and 5.8S rRNAs. The promoter of rRNA genes spans about 150 base pairs just upstream of the transcription initiation site (Cooper, 2000).

3.2.1.2 RNA polymerase III

The genes for tRNAs, 5S rRNA, and some of the small RNAs involved in splicing and protein transport are transcribed by RP III. These genes are characterized by promoters that lie within, rather than upstream of, the transcribed sequence. This is illustrated in Figure 3.1 (Cooper, 2000).

3.2.1.3 RNA Polymerase II

Because RP II is responsible for the synthesis of mRNA from protein-coding genes, it has been the focus of most studies of transcription in eukaryotes.

It consist of 12 subunits the largest of which contains a repeated heptapeptide sequence in its carboxyl-terminal domain (CTD), which is repeated 52 times in mouse, 42 times in Drosophila, 27 times in yeast, and is absent only in the *T. brucei* RPII subunit. The CTD structure is specific for the RP II subunit and is essential for cell viability. The CTD may function in initiation of transcription by mediating phosphorylation and destabilization of histone and DNA interactions, thus facilitating transcription through nucleosomes. Alternatively, the CTD may function as a receptor for essential TFs, or to anchor the RP II to a structure within the nucleus (Li *et al.*, 1989).

Little is known about transcription in *P. falciparum*, or its regulation. RNA synthesis in *P. falciparum*, during the 48 hour intra-erythrocytic growth cycle *in vitro*, starts within 6 hours after infection of RBCs, rapidly rises during the early trophozoite stage and peaks during the schizont stage. Many species of the parasite's proteins, including most of the known parasite antigens, are synthesized in synchronously infected cultures during the trophozoite and schizont stages (25−42 hours after infection of RBC) (Li *et al.*, 1989).

3.2.2 DISCOVERY

In the DISCOVERY database the following data is available:

3.2.2.1 Summary

The "Summary" tab provides the user with the protein identifier from PlasmoDB, in the case of PF11_0058-b: RNA polymerase subunit it is "PF11_0058-b". Known aliases of the protein is also provided and, in this case, it is only "RNA polymerase subunit, putative". The number of associated papers (seven), the protein sequence and the AC (Q8IIV5) is also given.

3.2.2.2 Function

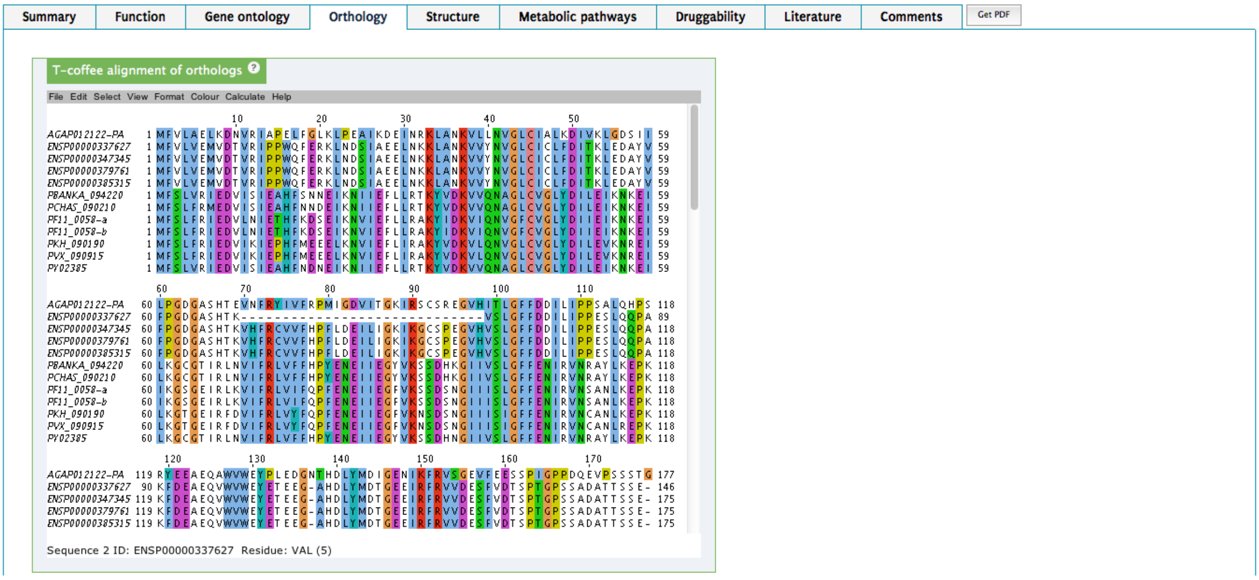
The "Function" tab revealed two different Interpro identifiers that matched the query, they were:

1. IPR013238, RNA polymerase III, subunit Rpc25. A strongly conserved subunit of RNA polymerase III that also have homology to Rpa43 in RNA polymerase 1, Rpb7 in RNA polymerase II and the archaral RpoE subunit. Rpc25 is required for transcription and initiation and is non-essential for the elongating properties of RNA polymerase III. And,
2. IPR005576, RNA polymerase Rpb7, N-terminal. Eukaryotic RNA polymerase subunits RPB4 and RPB7 form a heterodimer that reversibly associates with the RNA II core. Archaeal cells contain a single RP made up of 12 subunits, displaying considerable homology to the eukaryotic RPII subunits. RPB4 and RPB7 homologues are called subunits F and E, respectively, and it has been shown to form a stable heterodimer. While the RPB7 homologue is reasonably well conserved, the similarity between the eukaryotic RPB4 and the archaeal F subunit is barely detectable.

The InterPro entries can be summarized in Table 3.1.

Table 3.1: Summary of the InterPro signatures matching the RP subunit.

InterPro entry	Signatures	Analysis Method
IPR0123238 (Domain)	PF08292	HMMPfam
IPR005576(Domain)	PF03876	HMMPfam
Other	G3DSA:3.30.1490.120	Gene3D
	SSF88798	superfamily
	PTHR12709	HMMPanther
	PTHR12709:SF2	HMMPanther



```
                            10        20        30        40        50
AGAP012122-PA     1 MFVLAELKDNVRIAPELFGLKLPEAIKDEINRKLANKVLLNVGLCIALKDIVKLGDSII 59
ENSP00000337627   1 MFVLVEMVDTVRIPPWQFERKLNDSIAEELNKKLANKVVYNVGLCICLFDITKLEDAYV 59
ENSP00000347345   1 MFVLVEMVDTVRIPPWQFERKLNDSIAEELNKKLANKVVYNVGLCICLFDITKLEDAYV 59
ENSP00000379761   1 MFVLVEMVDTVRIPPWQFERKLNDSIAEELNKKLANKVVYNVGLCICLFDITKLEDAYV 59
ENSP00000385315   1 MFVLVEMVDTVRIPPWQFERKLNDSIAEELNKKLANKVVYNVGLCICLFDITKLEDAYV 59
PBANKA_094220     1 MFSLVRIEDVISIEAHFSNNEIKNIIEFLLRTKYVDKVVQNAGLCVGLYDILEIKNKEI 59
PCHAS_090210      1 MFSLFRMEDVISIEAHFNNDEIKNIIEFLLRTKYVDKVVQNAGLCVGLYDILEIKNKEI 59
PF11_0058-a       1 MFSLFRIEDVLNIETHFKDSEIKNVIEFLLRAKYIDKVIQNVGFCVGLYDIIEIKNKEI 59
PF11_0058-b       1 MFSLFRIEDVLNIETHFKDSEIKNVIEFLLRAKYIDKVIQNVGFCVGLYDIIEIKNKEI 59
PKH_090190        1 MFSLFRIEDVIKIEPHFMEEELKNVIEFLIRAKYVDKVLQNVGLCVGLYDILEVKNREI 59
PVX_090915        1 MFSLFRIEDVIKIEPHFMEEELKNVIEFLIRAKYVDKVLQNVGLCVGLYDILEVKNREI 59
PY02385           1 MFSLVRIEDVISIEAHFNDNEIKNIIEFLLRTKYVDKVVQNAGLCVGLYDILEIKNKEI 59

                    60        70        80        90        100       110
AGAP012122-PA    60 LPGDGASHTEVNFRYIVFRPMIGDVITGKIRSCSREGVHITLGFFDDILIPPSALQHPS 118
ENSP00000337627  60 FPGDGASHTK-----------------------------VSLGFFDDILIPPESLQQPA 89
ENSP00000347345  60 FPGDGASHTKVHFRCVVFHPFLDEILIGKIKGCSPEGVHVSLGFFDDILIPPESLQQPA 118
ENSP00000379761  60 FPGDGASHTKVHFRCVVFHPFLDEILIGKIKGCSPEGVHVSLGFFDDILIPPESLQQPA 118
ENSP00000385315  60 FPGDGASHTKVHFRCVVFHPFLDEILIGKIKGCSPEGVHVSLGFFDDILIPPESLQQPA 118
PBANKA_094220    60 LKGCGTIRLNVIFRLVFFHPYENEIIEGYVKSSDHKGIIVSLGFFENIRVNRAYLKEPK 118
PCHAS_090210     60 LKGCGTIRLNVIFRLVFFHPYENEIIEGYVKSSDHKGIIVSLGFFENIRVNRAYLKEPK 118
PF11_0058-a      60 IKGSGEIRLKVIFRLVIFQPFENEIIEGFVKSSDSNGIIISLGFFENIRVNSANLKEPK 118
PF11_0058-b      60 IKGSGEIRLKVIFRLVIFQPFENEIIEGFVKSSDSNGIIISLGFFENIRVNSANLKEPK 118
PKH_090190       60 LKGTGEIRFDVIFRLVYFQPFENEIIEGFVKNSDSNGIIISIGFFENIRVNCANLKEPK 118
PVX_090915       60 LKGTGEIRFDVIFRLVYFQPFENEIIEGFVKNSDSNGIIISIGFFENIRVNCANLREPK 118
PY02385          60 LKGCGTIRLNVIFRLVFFHPYENEIIEGYVKSSDHNGIIVSLGFFENIRVNRAYLKEPK 118

                      120       130       140       150       160       170
AGAP012122-PA   119 RYEEAEQAWVWEYPLEDGNTHDLYMDIGENIKFRVSGEVFEESSPIGPPDQEVPSSSTG 177
ENSP00000337627  90 KFDEAEQVWVWEYETEEG-AHDLYMDTGEEIRFRVVDESFVDTSPTGPSSADATTSSE- 146
ENSP00000347345 119 KFDEAEQVWVWEYETEEG-AHDLYMDTGEEIRFRVVDESFVDTSPTGPSSADATTSSE- 175
ENSP00000379761 119 KFDEAEQVWVWEYETEEG-AHDLYMDTGEEIRFRVVDESFVDTSPTGPSSADATTSSE- 175
ENSP00000385315 119 KFDEAEQVWVWEYETEEG-AHDLYMDTGEEIRFRVVDESFVDTSPTGPSSADATTSSE- 175
```

Sequence 2 ID: ENSP00000337627 Residue: VAL (5)

Figure 3.2: Screenshot of the T-coffee alignment of orthologs for PF11_0058b.

3.2.2.3 Gene Onthology

The "Gene Onthology" tab revealed two GO terms associated with this RP. They were:

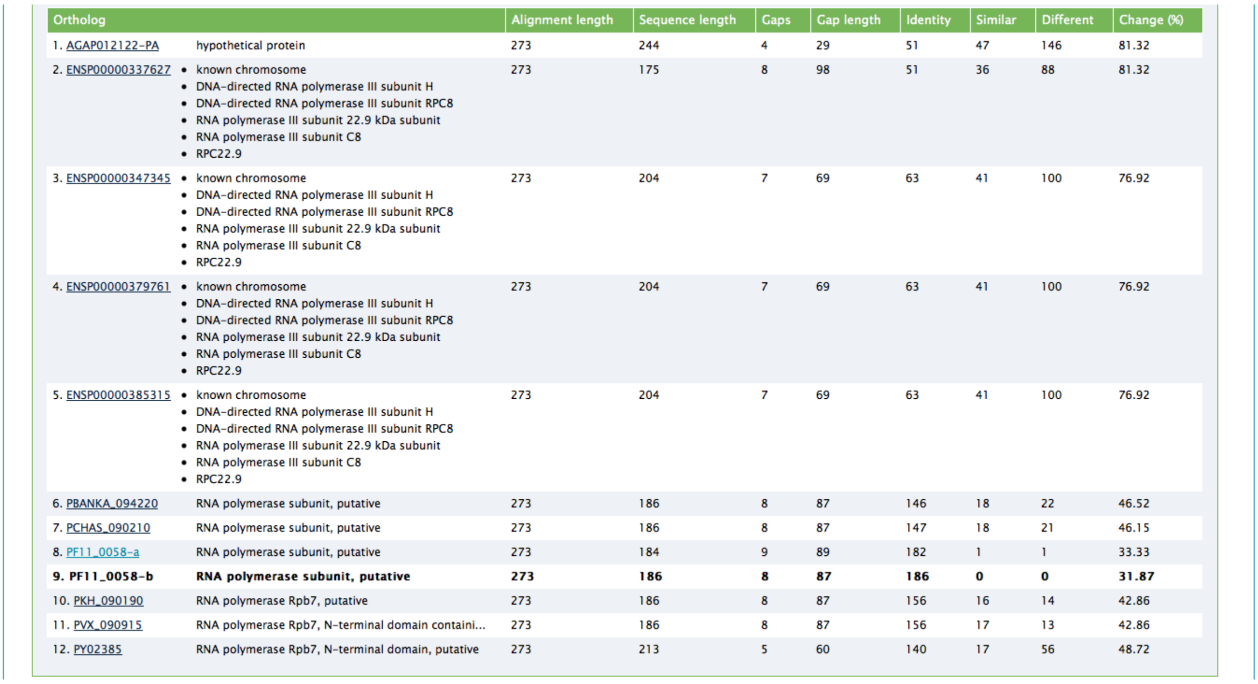
- GO:0003899, DNA-directed RNA polymerase activity. This GO term is associated at the molecular function level, with an evidence code of IEA and,
- GO:0006351, transcription, DNA-dependent. This GO term functions at the biological process level, with an evidence code of IEA.

No GO terms were found associated at the cellular component level.

3.2.2.4 Orthology

At the "Orthology" tab the OrthoMCL clustering can be found as well as an Ortholog table. The OrthoMCL clustering, Figure 3.2, showed that the RNA polymerase subunit is present in most of the species in DISCOVERY2 but is not very well conserved, and is a hypothetical protein in *Anopheles gambiae* and a putative protein in the majority of the *Plasmodium* species.

From the Ortholog table (see Figure 3.3) it was observed that all had an alignment length of 273 with *A. gambiae* having the longest sequence length at 244. The four *H. sapiens* sequences were roughly conserved with each other, each having 7 gaps with a gap length of 69, one had an extra gap, giving it a gap length of 98. The *Plasmodium* species were again slightly more conserved with each other, except for *P. yoelii* that had a longer sequence length and less gaps.



Ortholog		Alignment length	Sequence length	Gaps	Gap length	Identity	Similar	Different	Change (%)
1. AGAP012122-PA	hypothetical protein	273	244	4	29	51	47	146	81.32
2. ENSP00000337627	• known chromosome • DNA-directed RNA polymerase III subunit H • DNA-directed RNA polymerase III subunit RPC8 • RNA polymerase III subunit 22.9 kDa subunit • RNA polymerase III subunit C8 • RPC22.9	273	175	8	98	51	36	88	81.32
3. ENSP00000347345	• known chromosome • DNA-directed RNA polymerase III subunit H • DNA-directed RNA polymerase III subunit RPC8 • RNA polymerase III subunit 22.9 kDa subunit • RNA polymerase III subunit C8 • RPC22.9	273	204	7	69	63	41	100	76.92
4. ENSP00000379761	• known chromosome • DNA-directed RNA polymerase III subunit H • DNA-directed RNA polymerase III subunit RPC8 • RNA polymerase III subunit 22.9 kDa subunit • RNA polymerase III subunit C8 • RPC22.9	273	204	7	69	63	41	100	76.92
5. ENSP00000385315	• known chromosome • DNA-directed RNA polymerase III subunit H • DNA-directed RNA polymerase III subunit RPC8 • RNA polymerase III subunit 22.9 kDa subunit • RNA polymerase III subunit C8 • RPC22.9	273	204	7	69	63	41	100	76.92
6. PBANKA_094220	RNA polymerase subunit, putative	273	186	8	87	146	18	22	46.52
7. PCHAS_090210	RNA polymerase subunit, putative	273	186	8	87	147	18	21	46.15
8. PF11_0058-a	RNA polymerase subunit, putative	273	184	9	89	182	1	1	33.33
9. PF11_0058-b	**RNA polymerase subunit, putative**	**273**	**186**	**8**	**87**	**186**	**0**	**0**	**31.87**
10. PKH_090190	RNA polymerase Rpb7, putative	273	186	8	87	156	16	14	42.86
11. PVX_090915	RNA polymerase Rpb7, N-terminal domain containi...	273	186	8	87	156	17	13	42.86
12. PY02385	RNA polymerase Rpb7, N-terminal domain, putative	273	213	5	60	140	17	56	48.72

Figure 3.3: Screenshot of the ortholog table for PF11_0058b.

3.2.2.5 Structure

The "Structure" tab shows the two PDB BLAST sequences that have more that 70% coverage. If one clicks on the sequence link the alignment of the sequence with PF11_0058-b is shown. No ModBase structures were available for this protein.

3.2.2.6 Metabolic pathways

In the "Metabolic pathway" tab it is shown that the KEGG pathway associates this protein with the following pathways:

- Purine metabolism;
- Pyrimidine metabolism;
- Metabolic pathways and,
- RNA polymerase.

* These pathway links were collected when KEGG was still publicly available.

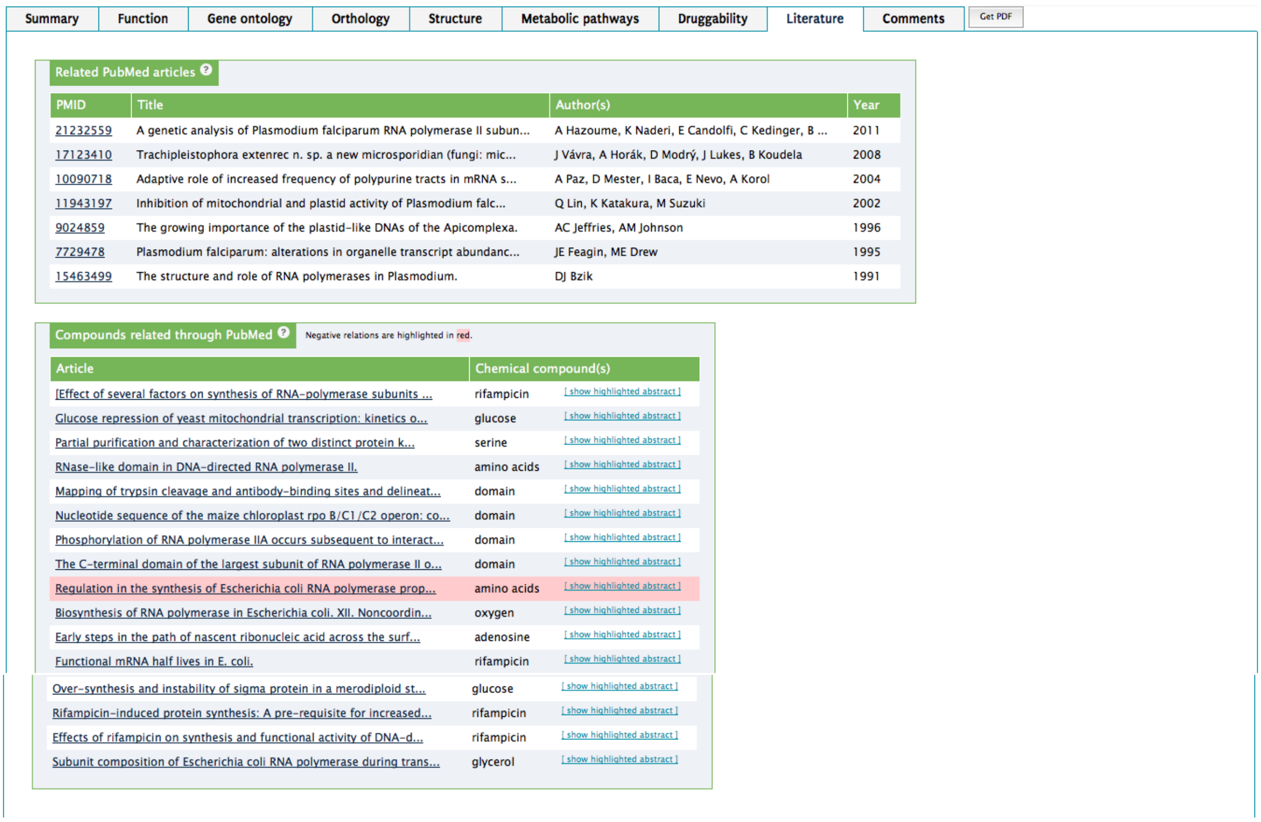
3.2.2.7 Druggability

The "Druggability" tab shows molecules with the most significant BLAST domain matches from DrugEBIlity and for this protein it is not predicted as druggable. There was only one gene that had a significant alignment to PF11_0058-b, namely P35718; and two domains had matches, both alignments were made against the domain 2CKZ.

3.2.2.8 Literature

The "Literature" tab is shown in Figure 3.4. Here it shows that there are seven articles associated with this RNA polymerase. These articles are retrieved from PubMed abstracts. Based on text relation mining of the abstracts the following are possible interactors of this RNA polymerase subunit:

- Rifampicin;



Related PubMed articles

PMID	Title	Author(s)	Year
21232559	A genetic analysis of Plasmodium falciparum RNA polymerase II subun...	A Hazoume, K Naderi, E Candolfi, C Kedinger, B ...	2011
17123410	Trachipleistophora extenrec n. sp. a new microsporidian (fungi: mic...	J Vávra, A Horák, D Modrý, J Lukes, B Koudela	2008
10090718	Adaptive role of increased frequency of polypurine tracts in mRNA s...	A Paz, D Mester, I Baca, E Nevo, A Korol	2004
11943197	Inhibition of mitochondrial and plastid activity of Plasmodium falc...	Q Lin, K Katakura, M Suzuki	2002
9024859	The growing importance of the plastid-like DNAs of the Apicomplexa.	AC Jeffries, AM Johnson	1996
7729478	Plasmodium falciparum: alterations in organelle transcript abundanc...	JE Feagin, ME Drew	1995
15463499	The structure and role of RNA polymerases in Plasmodium.	DJ Bzik	1991

Compounds related through PubMed

Negative relations are highlighted in red.

Article	Chemical compound(s)	
[Effect of several factors on synthesis of RNA-polymerase subunits ...	rifampicin	[show highlighted abstract]
Glucose repression of yeast mitochondrial transcription: kinetics o...	glucose	[show highlighted abstract]
Partial purification and characterization of two distinct protein k...	serine	[show highlighted abstract]
RNase-like domain in DNA-directed RNA polymerase II.	amino acids	[show highlighted abstract]
Mapping of trypsin cleavage and antibody-binding sites and delineat...	domain	[show highlighted abstract]
Nucleotide sequence of the maize chloroplast rpo B/C1/C2 operon: co...	domain	[show highlighted abstract]
Phosphorylation of RNA polymerase IIA occurs subsequent to interact...	domain	[show highlighted abstract]
The C-terminal domain of the largest subunit of RNA polymerase II o...	domain	[show highlighted abstract]
Regulation in the synthesis of Escherichia coli RNA polymerase prop...	amino acids	[show highlighted abstract]
Biosynthesis of RNA polymerase in Escherichia coli. XII. Noncoordin...	oxygen	[show highlighted abstract]
Early steps in the path of nascent ribonucleic acid across the surf...	adenosine	[show highlighted abstract]
Functional mRNA half lives in E. coli.	rifampicin	[show highlighted abstract]
Over-synthesis and instability of sigma protein in a merodiploid st...	glucose	[show highlighted abstract]
Rifampicin-induced protein synthesis: A pre-requisite for increased...	rifampicin	[show highlighted abstract]
Effects of rifampicin on synthesis and functional activity of DNA-d...	rifampicin	[show highlighted abstract]
Subunit composition of Escherichia coli RNA polymerase during trans...	glycerol	[show highlighted abstract]

Figure 3.4: Screenshot of the Literature tab for PF11_0058b.

- Glucose;
- Serine;
- Amino acids;
- Oxygen;
- Adenosine and,
- Glycerol.

Some amino acids might have a negative effect when interacting with the RNA polymerase subunit.

3.2.3 Discussion

In DISCOVERY2 there is not a lot of data available for PF11_0058-b: RNA polymerase subunit. It is not a protein that is relatively conserved across the species; there is no structure available and it is not considered druggable. When PF11_0058-b was put through Taverna and scored it scored as follows:

Chokepoint: 0

Interaction: 0

Ortholog: 4.7

Ligand: 0

ModBase: 0

Literature: 10

Druggability: 0.

This gave it a total weighted score of 0.16. Even if the weights were assigned differently it would never be able to score more than 0.35. Thus PF11_0058-b: RNA polymerase subunit, putative can, at this time, be considered a less than ideal drug target for future drug design regarding malaria. If more data becomes available for PF11_0058-b it might be reconsidered.

This protein was chosen as it had a very low score. After observing the amount of data available on DISCOVERY and the amount of literature available mentioning the protein it is understandable that it had a low score. Very few of the criteria identified to be of importance had a score, but still some data available. After observing the amount of data available on DISCOVERY and the amount of literature available mentioning the protein it is understandable that it had a low score. Very few of the criteria identified to be of importance had a score. This lack of data might be due to insufficient analysis or to no data available for that specific criterion, despite considerable effort.

Keeping its function as RNA polymerase in mind it can be considered a non-ideal drug target as it is a type of protein that is present in some form in virtually all living organisms, conserved to some degree across organisms.

3.2.4 Conclusion

PF11_0058-b: RNA polymerase subunit, putative is not considered a drug target, according to its score. It might change if more information became available and it gets rescored, but this is not considered likely as it is an RNA polymerase, a type of protein needed by all living organisms to function. Even though it is not a highly conserved protein its human ortholog score is closer to zero than it is to ten, which is not ideal.

3.3 PFL2280w: Serine/Threonine protein kinase, putative

3.3.1 Introduction

Protein kinases (PK) is one of the largest and most functionally diverse gene families that regulate cell functions. It not only adds phosphate groups to substrate proteins but it also directs the activity, localization and overall function of most proteins and cellular processes. PKs are particularly prominent in signal transduction and co-ordination of complex functions, like the cell cycle. Some proteins have multiple phosphorylation sites, this has a distinct or even opposing effects in the regulation of the protein (Manning, N/A, 2006).

Figure 3.5 illustrate protein phosphorylation and dephosphorylation. PKs bind substrate proteins and ATP (or ADP(P)); transferring a phosphate group from ATP to amino acids with free hydroxyl (-OH) groups, like serine, threonine or tyrosine. This results in a phospho-protein and ADP. Most PKs act on serine or threonine, but some are specific to tyrosine, while others might act on all three amino acids. The phosphate group on the protein (PO4-) is negatively charged and changes the substrate protein in different ways, for example; it alters the activity of the protein, including other kinases; the location of the protein and its turnover or interaction with other proteins. These changes are reversible by phosphatases, which remove the phosphate groups (Manning, 2006).

The diversity of essential functions mediated by PKs can be seen in the conservation of distinct kinase families between yeast, invertebrate and mammalian kinomes. Most mammalian PKs share a common structural domain, the eukaryotic PK domain. Kinases with this domain are almost exclusively eukaryotic, and constitute one of the largest eukaryotic gene families, representing $1 - 4\%$ of all genes in sequenced genomes. Comparison of the kinomes of several divergent genomes indicates that the most primitive eukaryote had about 35 distinct kinase functions (Manning, N/A, 2006).

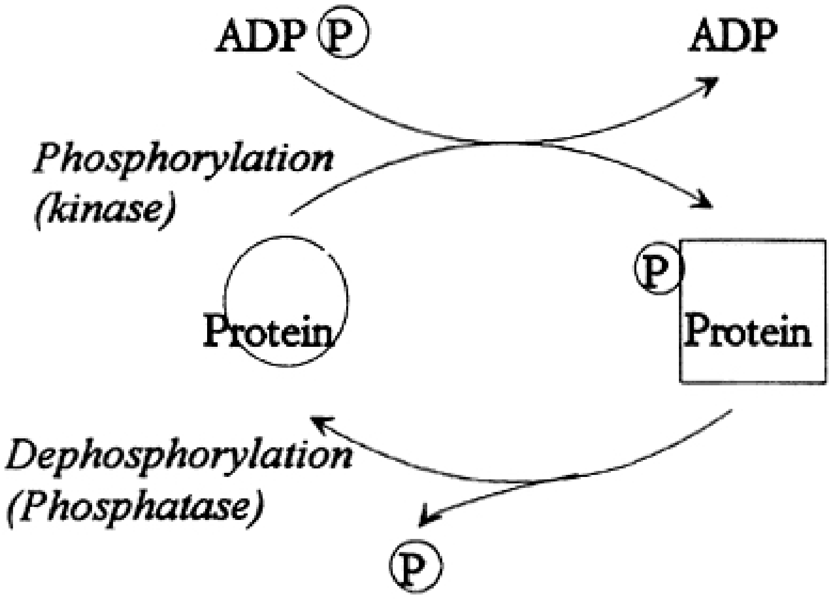


Figure 3.5: Reversible protein phosphorylation.
A PK moves a phosphate group (P) from ATP (ADP(P)) to the protein. The biological properties of the protein are thereby altered. There is also a protein phosphatase that is able to remove the phosphate group. The amount of phosphate that is associated with the protein is thus determined by the relative activities of the kinase and the phosphatase.
Source: http://www.nobelprize.org/nobel_prizes/medicine/laureates/1992/press.html.

3.3.1.1 Biological Functions

PKs can modify virtually all regulated biochemical pathways and complex behaviours. Some of the more common and more universal roles of PKs include:

- **Cell cycle control:** Cyclin dependent kinases control the various checkpoints that control cell cycle;
- **Response to extracellular stimuli:** Receptor kinases receive extracellular signals, and intracellular kinase cascades, such as the MAPK cascade, transduce this signal to various cellular components, including transcription;
- **DNA damage response:** The PIKK family is key mediators that perceive damaged DNA and co-ordinate the repair response and,
- **Metabolic control** (Manning, 2006).

3.3.1.2 Kinases and disease

Due to its key role in cell function and strong negative effects when misregulated PKs are considered an important protein in health and diseases. There are more than 160 PKs associated with human diseases and currently a major area of drug research is the inhibition of kinases (Manning, 2006). Several small molecule drugs and antibodies targeting kinases are already on the market, mostly as anticancer targets, and hundreds of kinase inhibitors are in various stages of development (Manning, 2006).

3.3.1.3 Other kinds of kinases

Histidine kinases

Histidine kinases phosphorylate themselves on histidine, before transferring that phosphate to an aspartate on a substrate protein. Histidine kinases are common in bacteria, plants and lower eukaryotes, but are not normally found in animals. They are structurally distinct from the usual form of serine/threonine/tyrosine kinases. Animals do have one class of kinases - mitochondrial pyruvate dehydrogenase kinases - that are structurally similar to histidine kinases, but it still phosphorylates on serine (Manning, N/A, 2006).

Tyrosine phosphorylation

Tyrosine phosphorylation has been of particular interest to many biologists, due to its biological roles. Most tyrosine phosphorylation is carried out by a distinct group of eukaryotic PK, called tyrosine kinase, though several serine/threonine-looking kinases also have tyrosine kinase abilities. Most tyrosine kinases are either receptor tyrosine kinases, whose extracellular region senses extracellular signals; or are receptor-associated kinases, that are located near the surface of the cell and interact with receptor tyrosine kinases (Manning, 2006).

There is some evidence for phosphorylation of other amino acids, though these phoshoamino acids are very liable, making them difficult to work with, and the responsible kinases have not yet been found (Manning, 2006). A wide variety of other kinases phosphorylate various small molecules, including those involved in metabolism, such as glycolysis or nucleotide metabolism. These come from a wide variety of different structures, distinct from that of PKs. One notable exception is the phosphatidyl inositol 3' kinases (PI3K). These are structurally related to PKs, and a subset of PI3K, known as PIKK, do phosphorylate proteins rather than small molecules (Manning, 2006).

3.3.1.4 Serine/Threonine protein kinase

Serine/Threonine PKs are the most common form of PK and can be defined as the following: "*a protein which catalyses the phosphorylation of serine or threonine residues on target proteins by using ATP as phosphate donor. Such phosphorylation may cause changes in the function of the target protein.*" (UniProtFAQ, N/A).

3.3.2 DISCOVERY

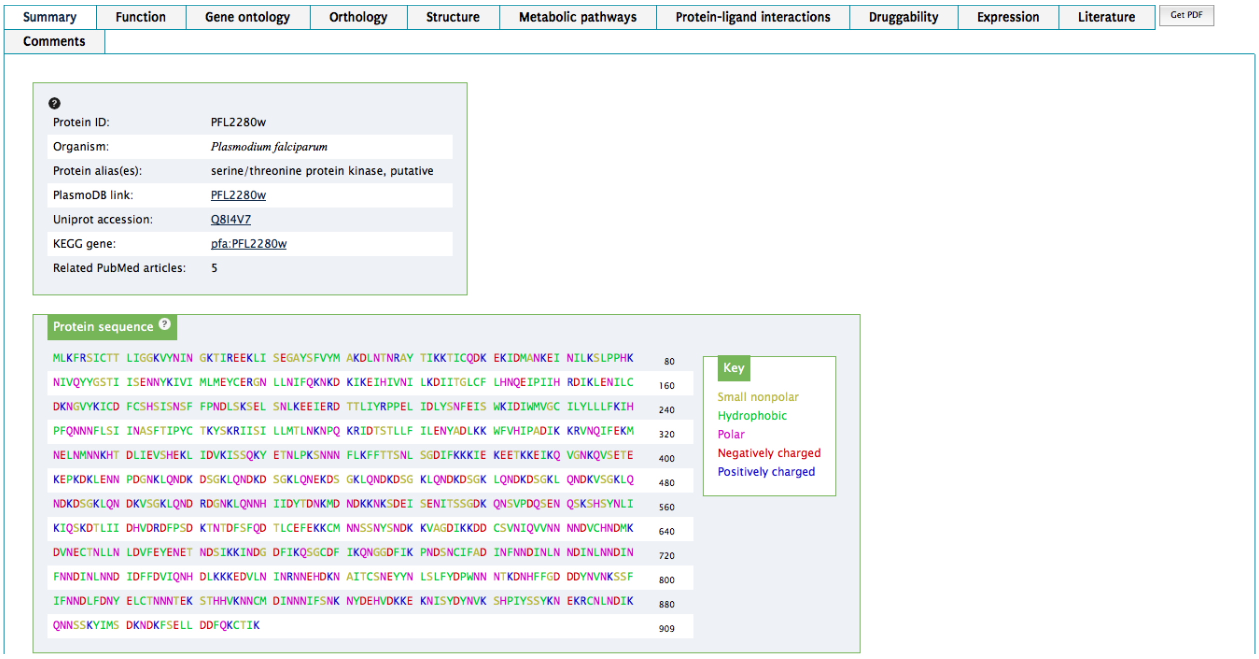
3.3.2.1 Summary

The "Summary" tab, as can be seen in Figure 3.6 and provides the user with the PlasmoDB identifier, and in the case of PFL2280w: serine/threonine PK it will be "PFL2280w". The known alias of this protein is only "serine/threonine PK, putative". The number of associated papers is five; and the protein sequence is also given as well as the Uniprot accession number, which is Q8I4V7 for PFL2280w: serine/threonine PK, putative.

3.3.2.2 Function

The "Function" tab revealed four different Interpro identifiers that matched the query, PFL2280w, they were:

1. IPR008271, serine/threonine PK, active site. Protein phosphorylation plays a key role in most cellular activities, it is a reversible process mediated by PKs and phospho-protein phosphatases. PKs catalyse the transfer of the gamma phosphate from nucleotide triphosphates to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function. Phospho-protein phosphatases catalyse the reverse process. PKs fall



Protein ID:	PFL2280w
Organism:	*Plasmodium falciparum*
Protein alias(es):	serine/threonine protein kinase, putative
PlasmoDB link:	PFL2280w
Uniprot accession:	Q8I4V7
KEGG gene:	pfa:PFL2280w
Related PubMed articles:	5

Protein sequence

```
MLKFRSICTT LIGGKVYNIN GKTIREEKLI SEGAYSFVYM AKDLNTNRAY TIKKTICQDK EKIDMANKEI NILKSLPPHK   80
NIVQYYGSTI ISENNYKIVI MLMEYCERGN LLNIFQKNKD KIKEIHIVNI LKDIITGLCF LHNQEIPIIH RDIKLENILC  160
DKNGVYKICD FCSHSISNSF FPNDLSKSEL SNLKEEIERD TTLIYRPPEL IDLYSNFEIS WKIDIWMVGC ILYLLLFKIH  240
PFQNNNFLSI INASFTIPYC TKYSKRIISI LLMTLNKNPQ KRIDTSTLLF ILENYADLKK WFVHIPADIK KRVNQIFEKM  320
NELNMNNKHT DLIEVSHEKL IDVKISSQKY ETNLPKSNNN FLKFFTTSNL SGDIFKKKIE KEETKKEIKQ VGNKQVSETE  400
KEPKDKLENN PDGNKLQNDK DSGKLQNDKD SGKLQNEKDS GKLQNDKDSG KLQNDKDSGK LQNDKDSGKL QNDKVSGKLQ  480
NDKDSGKLQN DKVSGKLQND RDGNKLQNNH IIDYTDNKMD NDKKNKSDEI SENITSSGDK QNSVPDQSEN QSKSHSYNLI  560
KIQSKDTLII DHVDRDFPSD KTNTDFSFQD TLCEFEKKCM NNSSNYSNDK KVAGDIKKDD CSVNIQVVNN NNDVCHNDMK  640
DVNECTNLLN LDVFEYENET NDSIKKINDG DFIKQSGCDF IKQNGGDFIK PNDSNCIFAD INFNNDINLN NDINLNNDIN  720
FNNDINLNND IDFFDVIQNH DLKKKEDVLN INRNNEHDKN AITCSNEYYN LSLFYDPWNN NTKDNHFFGD DDYNVNKSSF  800
IFNNDLFDNY ELCTNNNTEK STHHVKNNCM DINNNIFSNK NYDEHVDKKE KNISYDYNVK SHPIYSSYKN EKRCNLNDIK  880
QNNSSKYIMS DKNDKFSELL DDFQKCTIK                                                         909
```

Key

Small nonpolar
Hydrophobic
Polar
Negatively charged
Positively charged

Figure 3.6: Screenshot of the Summary tab for PFL2280w.

into three broad classes, characterised with respect to substrate specificity:

(a) Serine/Threonine PKs;
(b) Tyrosine PKs and,
(c) Dual specificity PKs (e.g. MEK - phosphorylates both threonine and tyrosine on target proteins).

PK function is evolutionarily conserved from *Escherichia coli* to human. PKs play a role in a multitude of cellular processes, including division, proliferation, apoptosis, and differentiation. Phosphorylation usually results in a functional change of the target protein by changing enzyme activity, cellular location, or association with other proteins. The catalytic subunits of PKs are highly conserved, and several structures have been solved, leading to large screens to develop kinase-specific inhibitors for the treatments of a number of diseases.

Eukaryotic PK are enzymes that belong to a very extensive family of proteins, which share a conserved catalytic core common with both serine/threonine and tyrosine PKs. There are a number of conserved regions in the catalytic domain of PKs. In the N-terminal extremity of the catalytic domain there is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP binding. In the central part of the catalytic domain there is a conserved aspartic acid residue, which is important for the catalytic activity of the enzyme. This signature contains the active site aspartate residue;

2. IPR000719, PK domain. This entry represents the PK domain containing the catalytic function of PKs. This domain is found in serine/threonine PKs, tyrosine PKs and dual specificity PKs;
3. IPR011009, PK-like domain. PKs modify other proteins by chemically adding phosphate groups to them. This process is fundamental to most signalling and regulatory processes in the eukaryotic cell. The PKs contain a catalytic core that is common to both serine/threonine and tyrosine PKs. The catalytic domain contains the nucleotide-binding site and the catalytic apparatus in an inter-lobe cleft. Structurally it shares functional and structural similarities with the ATP-grasp fold,

Table 3.2: Summary of the InterPro signatures matching to serine/threonine PK.

InterPro entry	Signatures	Analysis Method
IPR008271 (Active site)	PS00108	PatternScan
IPR00719 (Domain)	PS5001	ProfileScan
IPR011009 (Domain)	SSF56112	superfamily
IPR017442 (Domain)	PF00069	HMMPfam
Other	PTHR22967:SF29	HMMPanther
	PTHR22967	HMMPanther
	G3DSA:1.10.510.10	Gene3D

which is found in enzymes that catalyse the formation of an amide bond. The 3D fold of the PK catalytic domain is similar to domains found in several other proteins. These include:

(a) The catalytic domain of PI3K, which phosphorylates phospho-inositides and, as such, is involved in a number of fundamental cellular processes such as apoptosis, proliferation, motility and adhesion;

(b) Choline kinase, which catalyses the ATP-dependent phosphorylation of choline during the biosynthesis of phosphatidylcholine and,

(c) 3',5'-aminoglycoside phosphotransferase type IIIa, a bacterial enzyme that confers resistance to a range of aminoglycoside antibiotics. And,

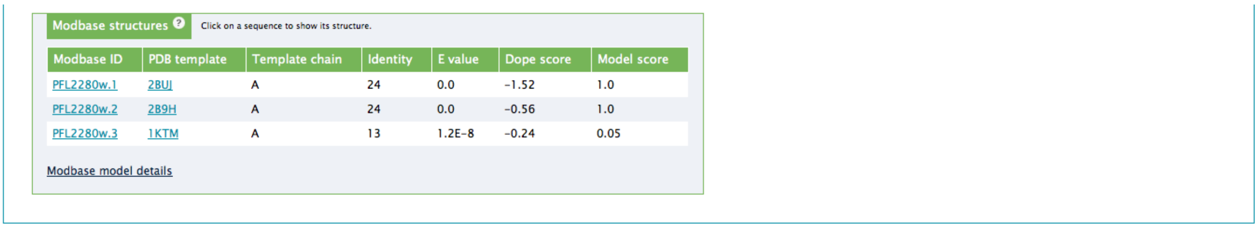
4. IPR017442 (have been changed to IPR000719), see point 2.

The InterPro entries can be summarized in Table 3.2.

3.3.2.3 Gene ontology

The "Gene Ontology" tab revealed multiple GO terms associated with the serine/threonine PK. They where:

- At the Cellular component level:
  - GO:0030140, trans-Golgi network transport vesicle, with an evidence code of IEA;
- At the Molecular function level:
  - GO:0004672, PK activity, with an evidence code of IEA;
  - GO:0004674, protein serine/threonine kinase activity, with an evidence code of IEA;
  - GO:0004693, cyclin-dependent PK activity, with an evidence code of IEA;
  - GO:0005524, ATP binding, with an evidence code of IEA;
  - GO:0016301, kinase activity, with an evidence code of IEA;
  - GO:0016740, transferase activity, with an evidence code of IEA and,
  - GO:0016772, transferase activity, transferring phosphorus-containing groups, with an evidence code of IEA. And,
- At the Biological processes level:
  - GO:0006468, protein phosphorylation, with an evidence code of IEA;
  - GO:0007049, cell cycle, with an evidence code of IEA and,
  - GO:0016310, phosphorylation, with an evidence code of IEA.



Modbase structures ? Click on a sequence to show its structure.

Modbase ID	PDB template	Template chain	Identity	E value	Dope score	Model score
PFL2280w.1	2BUJ	A	24	0.0	-1.52	1.0
PFL2280w.2	2B9H	A	24	0.0	-0.56	1.0
PFL2280w.3	1KTM	A	13	1.2E-8	-0.24	0.05

Modbase model details

Figure 3.7: Screenshot of the ModBase structures under the Structure tab for PFL2280w.

3.3.2.4 Orthology

At the "Orthology" tab the OrthoMCl clustering can be found as well as an Ortholog table. The OrthoMCL clustering showed that the serine/threonine protein kinase is present in most of the malaria species in DISCOVERY2. In all species it is a putative protein, except in *P. yoelii* it is a hypothetical protein. The amount of similar or identical amino acids between the species for this specific protein might not be ideal, but according to the alignment there are no human or *A. gambiae* orthologs.

From the Ortholog table it was observed that all had an alignment length of 979, except *P. yoelii* had a length of 623. All six proteins further had different sequence lengths, amount of gaps, gap lengths etc.

3.3.2.5 Structure

The "Structure" tab shows multiple sequences that had a 70% coverage using PDB BLAST. If one clicks on the sequence link the alignment of the sequence with PFL2280w is shown. There are three possible ModBase structures available as can be seen in Figure 3.7 for PFL2280w: serine/threonine protein kinase. They are:

- PFL2280w.1, using PDB template 2BUJ and has a DOPE score of -1.52;
- PFL2280w.2, using PDB template 2B9H and has a DOPE score of -0.56 and,
- PFL2280w.3 using PDB template 1KTM and has a DOPE score of -0.24.

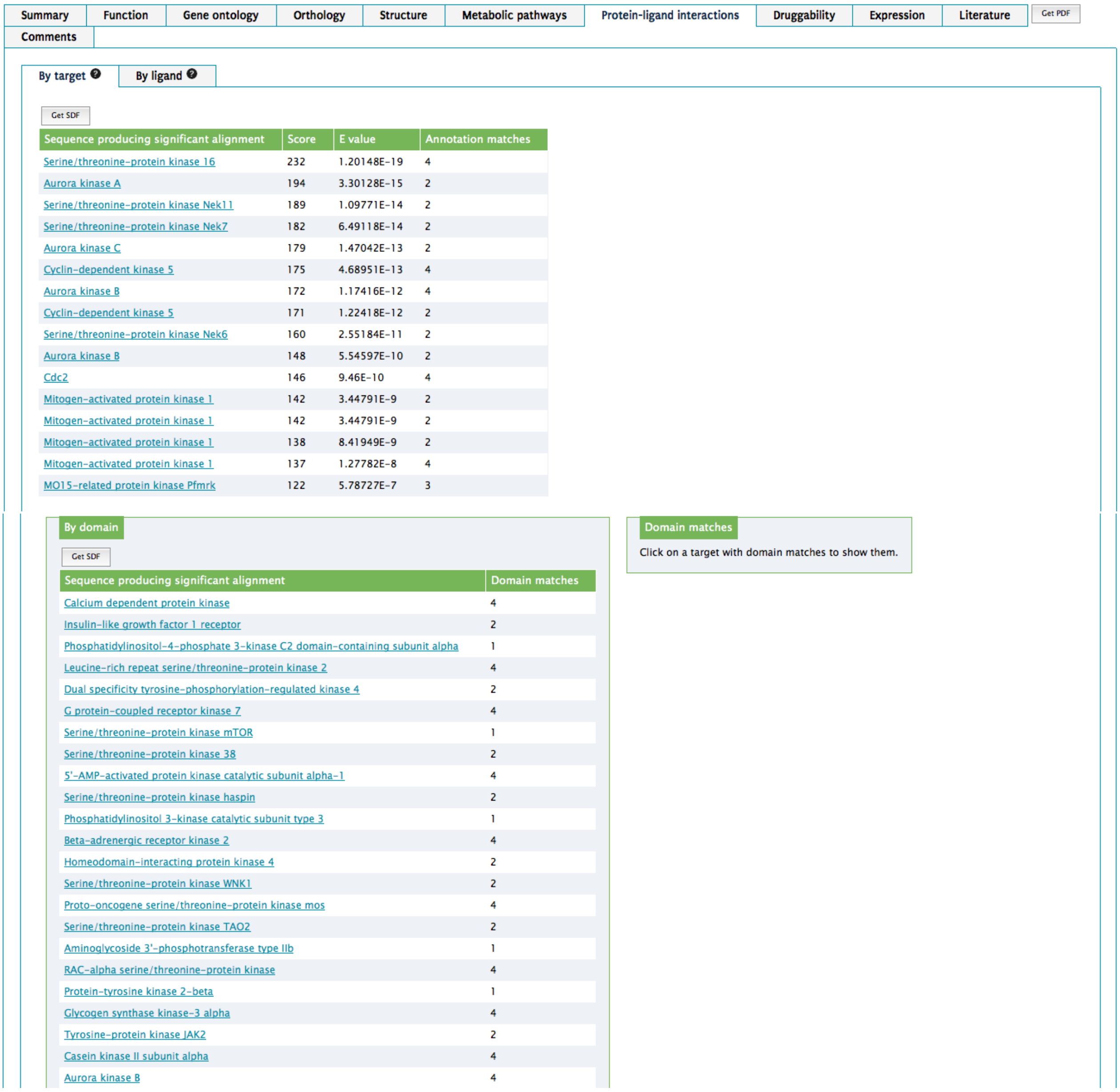
3.3.2.6 Metabolic pathways

In the "Metabolic pathway" tab it is shown that the MPMP pathways associates with this protein as follows:

- Protein kinase coding genes with activity serine/threonine protein kinase, putative;
- The phosphoproteome of *Plasmodium falciparum* infected RBCs with activity of serine/threonine protein kinase;
- Protein phosphorylation with activity of protein kinase and,
- Pre-replicative complex formation and transition to replication with activity of cyclindependent kinase.

The "Metabolic pathway" tab also supplies enzyme information in the form of EC numbers. For PFL2280w: serine/threonine protein kinase they are as follows:

- 2.7.11.1 Non-specific serine/threonine protein kinase and,
- 3.6.5.2 Small monomeric GTPase.



Sequence producing significant alignment	Score	E value	Annotation matches
Serine/threonine-protein kinase 16	232	1.20148E-19	4
Aurora kinase A	194	3.30128E-15	2
Serine/threonine-protein kinase Nek11	189	1.09771E-14	2
Serine/threonine-protein kinase Nek7	182	6.49118E-14	2
Aurora kinase C	179	1.47042E-13	2
Cyclin-dependent kinase 5	175	4.68951E-13	4
Aurora kinase B	172	1.17416E-12	4
Cyclin-dependent kinase 5	171	1.22418E-12	2
Serine/threonine-protein kinase Nek6	160	2.55184E-11	2
Aurora kinase B	148	5.54597E-10	2
Cdc2	146	9.46E-10	4
Mitogen-activated protein kinase 1	142	3.44791E-9	2
Mitogen-activated protein kinase 1	142	3.44791E-9	2
Mitogen-activated protein kinase 1	138	8.41949E-9	2
Mitogen-activated protein kinase 1	137	1.27782E-8	4
MO15-related protein kinase Pfmrk	122	5.78727E-7	3

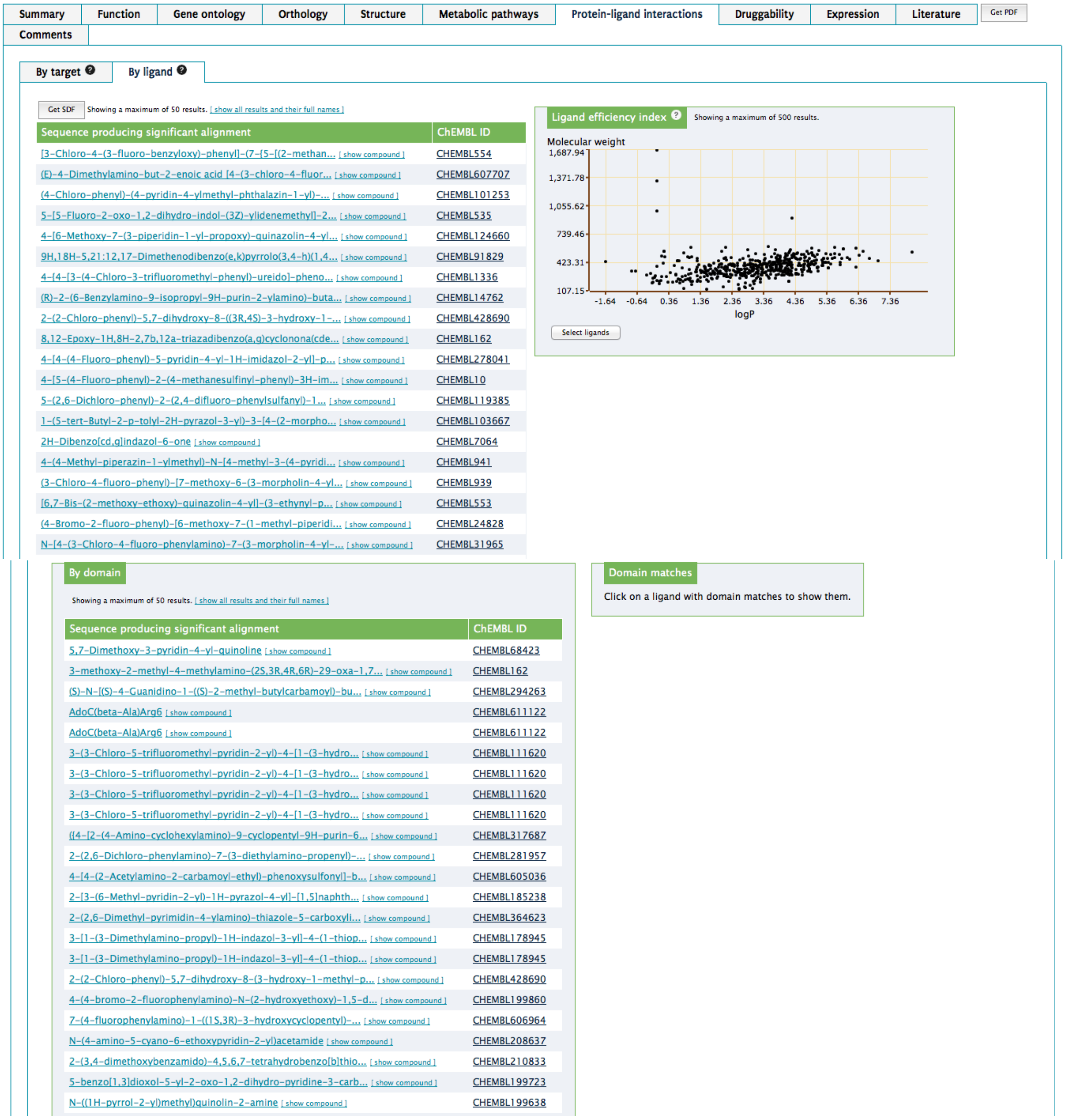
By domain

Sequence producing significant alignment	Domain matches
Calcium dependent protein kinase	4
Insulin-like growth factor 1 receptor	2
Phosphatidylinositol-4-phosphate 3-kinase C2 domain-containing subunit alpha	1
Leucine-rich repeat serine/threonine-protein kinase 2	4
Dual specificity tyrosine-phosphorylation-regulated kinase 4	2
G protein-coupled receptor kinase 7	4
Serine/threonine-protein kinase mTOR	1
Serine/threonine-protein kinase 38	2
5'-AMP-activated protein kinase catalytic subunit alpha-1	4
Serine/threonine-protein kinase haspin	2
Phosphatidylinositol 3-kinase catalytic subunit type 3	1
Beta-adrenergic receptor kinase 2	4
Homeodomain-interacting protein kinase 4	2
Serine/threonine-protein kinase WNK1	2
Proto-oncogene serine/threonine-protein kinase mos	4
Serine/threonine-protein kinase TAO2	2
Aminoglycoside 3'-phosphotransferase type IIb	1
RAC-alpha serine/threonine-protein kinase	4
Protein-tyrosine kinase 2-beta	1
Glycogen synthase kinase-3 alpha	4
Tyrosine-protein kinase JAK2	2
Casein kinase II subunit alpha	4
Aurora kinase B	4

Domain matches

Click on a target with domain matches to show them.

Figure 3.8: Screenshot of the Protein-ligand interaction tab, sub tab by target for PFL2280w.



Showing a maximum of 50 results. [show all results and their full names]

Sequence producing significant alignment	ChEMBL ID
[3-Chloro-4-(3-fluoro-benzyloxy)-phenyl]-(7-{5-[(2-methan... [show compound]	CHEMBL554
(E)-4-Dimethylamino-but-2-enoic acid [4-(3-chloro-4-fluor... [show compound]	CHEMBL607707
(4-Chloro-phenyl)-(4-pyridin-4-ylmethyl-phthalazin-1-yl)-... [show compound]	CHEMBL101253
5-[5-Fluoro-2-oxo-1,2-dihydro-indol-(3Z)-ylidenemethyl]-2... [show compound]	CHEMBL535
4-[6-Methoxy-7-(3-piperidin-1-yl-propoxy)-quinazolin-4-yl... [show compound]	CHEMBL124660
9H,18H-5,21:12,17-Dimethenodibenzo(e,k)pyrrolo(3,4-h)(1,4... [show compound]	CHEMBL91829
4-{4-[3-(4-Chloro-3-trifluoromethyl-phenyl)-ureido]-pheno... [show compound]	CHEMBL1336
(R)-2-(6-Benzylamino-9-isopropyl-9H-purin-2-ylamino)-buta... [show compound]	CHEMBL14762
2-(2-Chloro-phenyl)-5,7-dihydroxy-8-((3R,4S)-3-hydroxy-1-... [show compound]	CHEMBL428690
8,12-Epoxy-1H,8H-2,7b,12a-triazadibenzo(a,g)cyclonona(cde... [show compound]	CHEMBL162
4-[4-(4-Fluoro-phenyl)-5-pyridin-4-yl-1H-imidazol-2-yl]-p... [show compound]	CHEMBL278041
4-[5-(4-Fluoro-phenyl)-2-(4-methanesulfinyl-phenyl)-3H-im... [show compound]	CHEMBL10
5-(2,6-Dichloro-phenyl)-2-(2,4-difluoro-phenylsulfanyl)-1... [show compound]	CHEMBL119385
1-(5-tert-Butyl-2-p-tolyl-2H-pyrazol-3-yl)-3-[4-(2-morpho... [show compound]	CHEMBL103667
2H-Dibenzo[cd,g]indazol-6-one [show compound]	CHEMBL7064
4-(4-Methyl-piperazin-1-ylmethyl)-N-[4-methyl-3-(4-pyridi... [show compound]	CHEMBL941
(3-Chloro-4-fluoro-phenyl)-[7-methoxy-6-(3-morpholin-4-yl... [show compound]	CHEMBL939
[6,7-Bis-(2-methoxy-ethoxy)-quinazolin-4-yl]-(3-ethynyl-p... [show compound]	CHEMBL553
(4-Bromo-2-fluoro-phenyl)-[6-methoxy-7-(1-methyl-piperidi... [show compound]	CHEMBL24828
N-[4-(3-Chloro-4-fluoro-phenylamino)-7-(3-morpholin-4-yl-... [show compound]	CHEMBL31965

Showing a maximum of 500 results.

By domain

Showing a maximum of 50 results. [show all results and their full names]

Sequence producing significant alignment	ChEMBL ID
5,7-Dimethoxy-3-pyridin-4-yl-quinoline [show compound]	CHEMBL68423
3-methoxy-2-methyl-4-methylamino-(2S,3R,4R,6R)-29-oxa-1,7... [show compound]	CHEMBL162
(S)-N-[(S)-4-Guanidino-1-((S)-2-methyl-butylcarbamoyl)-bu... [show compound]	CHEMBL294263
AdoC(beta-Ala)Arg6 [show compound]	CHEMBL611122
AdoC(beta-Ala)Arg6 [show compound]	CHEMBL611122
3-(3-Chloro-5-trifluoromethyl-pyridin-2-yl)-4-[1-(3-hydro... [show compound]	CHEMBL111620
3-(3-Chloro-5-trifluoromethyl-pyridin-2-yl)-4-[1-(3-hydro... [show compound]	CHEMBL111620
3-(3-Chloro-5-trifluoromethyl-pyridin-2-yl)-4-[1-(3-hydro... [show compound]	CHEMBL111620
3-(3-Chloro-5-trifluoromethyl-pyridin-2-yl)-4-[1-(3-hydro... [show compound]	CHEMBL111620
({4-[2-(4-Amino-cyclohexylamino)-9-cyclopentyl-9H-purin-6... [show compound]	CHEMBL317687
2-(2,6-Dichloro-phenylamino)-7-(3-diethylamino-propenyl)-... [show compound]	CHEMBL281957
4-[4-(2-Acetylamino-2-carbamoyl-ethyl)-phenoxysulfonyl]-b... [show compound]	CHEMBL605036
2-[3-(6-Methyl-pyridin-2-yl)-1H-pyrazol-4-yl]-[1,5]naphth... [show compound]	CHEMBL185238
2-(2,6-Dimethyl-pyrimidin-4-ylamino)-thiazole-5-carboxyli... [show compound]	CHEMBL364623
3-[1-(3-Dimethylamino-propyl)-1H-indazol-3-yl]-4-(1-thiop... [show compound]	CHEMBL178945
3-[1-(3-Dimethylamino-propyl)-1H-indazol-3-yl]-4-(1-thiop... [show compound]	CHEMBL178945
2-(2-Chloro-phenyl)-5,7-dihydroxy-8-(3-hydroxy-1-methyl-p... [show compound]	CHEMBL428690
4-(4-bromo-2-fluorophenylamino)-N-(2-hydroxyethoxy)-1,5-d... [show compound]	CHEMBL199860
7-(4-fluorophenylamino)-1-((1S,3R)-3-hydroxycyclopentyl)-... [show compound]	CHEMBL606964
N-(4-amino-5-cyano-6-ethoxypyridin-2-yl)acetamide [show compound]	CHEMBL208637
2-(3,4-dimethoxybenzamido)-4,5,6,7-tetrahydrobenzo[b]thio... [show compound]	CHEMBL210833
5-benzo[1,3]dioxol-5-yl-2-oxo-1,2-dihydro-pyridine-3-carb... [show compound]	CHEMBL199723
N-((1H-pyrrol-2-yl)methyl)quinolin-2-amine [show compound]	CHEMBL199638

Domain matches

Click on a ligand with domain matches to show them.

Figure 3.9: Screenshot of the Protein-ligand interaction tab, sub tab by ligand for PFL2280w.

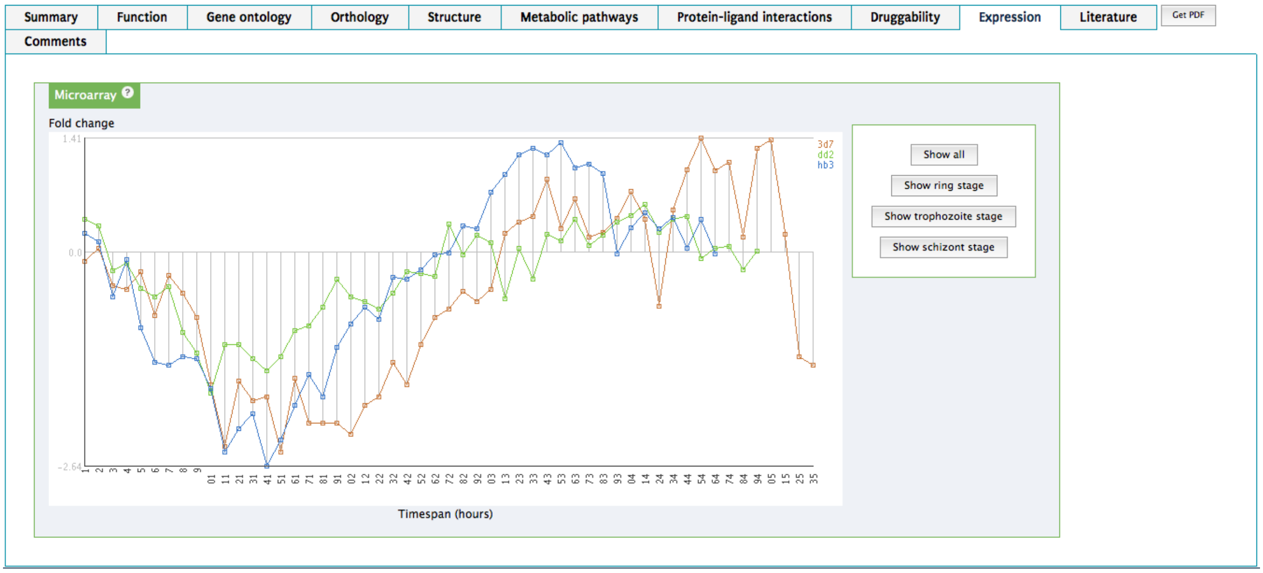


Figure 3.10: Screenshot of the Expression tab for PFL2280w.

3.3.2.7 Protein-ligand interactions

The "Protein-ligand interaction" tab can be seen in Figure 3.8 and Figure 3.9. Figure 3.8 depicts the "By target" sub tab while Figure 3.9 depict the "By ligand" sub tab. Using a BLAST search against the ChEMBL targets database identified the interactions reported in the targets section; the associated ligands are reported in the ligand section.

The target database (Figure 3.8) reveals a few homologs, identified by BLAST. They include mostly human homologs. Domain matched sequence include proteins that are kinaselike and include the proteins already matched according to the full sequence, it also includes insulin-receptors and insulin-like growth factor 1 receptor.

The ligand database (Figure 3.9) matched a few thousand chemical compounds that have been tested against PFL2280w: serine/threonine protein kinase, putative and its homologs. The plot in the top right hand corner of Figure 3.9 plots ligand efficiency indexes, graphing logP against molecular weight for every ligand. Ligand efficiency is a measurement of the binding energy of a ligand to a binding partner, like a receptor or enzyme.

3.3.2.8 Druggability

The "Druggability" tab shows molecules with the most significant BLAST domain matches from DrugEBIlity and for this serine/threonine protein kinase it is predicted as druggable. PFL2280w has multiple domains, identified by BLAST that can be considered druggable. There are also multiple genes, identified by BLAST comparisons, with viable scores and E values to consider PFL2280w: serine/threonine protein kinase as a potential drug target.

3.3.2.9 Expression

In the "Expression" tab it seems that PFL2280w: serine/threonine PK is predominantly expressed in the schizont stage and proposes that it is under expressed in the early ring stage. This can be seen in Figure 3.10 and may indicate that PFL2280w: serine/threonine PK plays a role in the asexual reproduction, or the division of the parasite cells. If this conclusion is correct it can be a viable drug target.

3.3.2.10 Literature

The “Literature” tab showed that there are five articles associated with this serine/threonine PK. These articles are retrieved from PubMed abstracts. Based on text relation mining of the abstracts it is possible that the following might be interactors of this kinase:

- Tyrosine;
- Proline;
- Serine and Threonine;
- Glycerol;
- Propofol;
- Lysine;
- Dopamine and Glutathione;
- Adenosine;
- Rapamycin;
- Inositol and,
- Proline and Penicillin.

3.3.3 Discussion

In DISCOVERY2 there is more data available for PFL2280w: serine/threonine PK, putative than there was for PF11_0058-b: RNA polymerase subunit; it is not a protein that is ideally conserved across the *Plasmodium* genus, but there is also no ortholog in humans and mosquitos; there is a crystal structure available, possible ligands to bind to this target and this kinase is considered druggable. Experimental protein-ligand interactions and expression data was also found for this kinase by a single search on DISCOVERY2. When PFL2280w was put through Taverna it scored as follows:

Chokepoint:	0
Interaction:	2.2
Ortholog:	10
Ligand:	4.54
ModBase:	10
Literature:	10
Druggability:	10.

This gives it a total weighted score of 0.5. If the weights were assigned differently the maximum weighted score PFL2280w would be able to achieve is 0.85 and the minimum would be 0.49. Thus PFL2280w: serine/threonine PK, putative can be considered a possible drug target for future drug design regarding malaria, depending on what is considered important for the drug company. The fact that it is not considered as a chokepoint weighs the protein down.

PFL2280w was chosen as a protein with an intermittent score. With a score 0.5 the protein can be considered a plausible drug target if chokepoints are not considered an important criterion, or if it is identified as a chokepoint in the future. Using DISCOVERY, more information was available, when compared to PF11_0058-b. This alone can be promising, but as most data for all the criteria identified has been observed it increases the viability of this protein as a potential drug target. Some of the lower scored criteria received full marks with the higher scored criteria receiving less than perfect scores. Due to this, if the importance of criteria is rearranged the protein can be considered very important in future drug design.

3.3.4 Conclusion

PFL2280w: serine/threonine PK, putative is not considered a drug target, according to its score. If the scores are rearranged, and "Chokepoint" is not considered a criterion (or this protein is identified as a chokepoint in the future) it might be considered as a drug target in *P. falciparum.* If expression data is taken into consideration and forms part of the score PFL2280w: serine/threonine PK, putative might score higher.

3.4 PFE1050w: S-adenosyl-L-homocysteine hydrolase

3.4.1 Introduction

The reversible conversion of S-adenosyl-L-homocysteine (SAH) to adenosine and homocysteine involves a unique enzyme, S-adenosyl-L-homocysteine hydrolase (SAHH) (Khare *et al.*, 2012; Malanovic *et al.*, 2008; Nakanishi, 2007). SAHH is an enzyme that is highly conserved from bacteria to mammals (Malanovic *et al.*, 2008) and occurs downstream of the S-adenosyl-methionine (SAM) dependent transmethylation enzymes, which are responsible for a wide variety of important biological functions, like gene expression, signal transduction, and lipid biosynthesis (Malanovic *et al.*, 2008). Since, SAH is a product inhibitor of all SAM-dependent methyltransferases, the catalytic activity of SAHH is critical in eukaryotic cells to maintain the normal cellular level of SAH and to permit the numerous transmethylation reactions required in normal cell functions to proceed. Because of the essential roles of SAHH for the living cells, inhibition of SAHH in eukaryotic cells causes an increase in the cellular level of SAH and inhibition of SAM-dependent methyltransferases. This results in various pharmacological effects that may include: antiviral, antiparasitic, antiarthritic and immunosuppressive (Khare *et al.*, 2012; Nakanishi, 2007).

3.4.1.1 Mechanism

The hydrolysis of SAH via SAHH is NAD^+ dependent and loosely involves two steps. First the NAD^+ binds with the SAHH at the NAD^+ binding site, and thereafter, the enzyme hydrolyses the substrate SAH into adenosine and homocysteine (Khare *et al.*, 2012). But, the hydrolysis mechanism from SAH to adenosine and homocysteine can be better described by the following six steps that is also illustrated in Figure 3.11 (indicated by numbers one to six):

1. In the initial step of SAHH the $N\zeta$ atom of lysine235 accepts the proton from the 3´-OH group of the substrate;
2. The nucleophilic character of lysine235 is increased by hydrogen bonds to the $O\delta1$ atoms of asparagine230 and asparagine240 and to the carbonyl-O-atom of glutamic acid205. As a result, the C3´–H H-atom becomes more liable and hydride abstraction by the NAD^+ cofactor is facilitated. The products of this reaction step are 3´-keto-adenosinehomocysteine and NADH;
3. The acidity of the C4´—H group of the 3´-keto derivative is higher when compared with the adenosine molecule, allowing the formation of a C4´– carbo-anion through proton transfer to the carboxylic group of aspartic acid139;

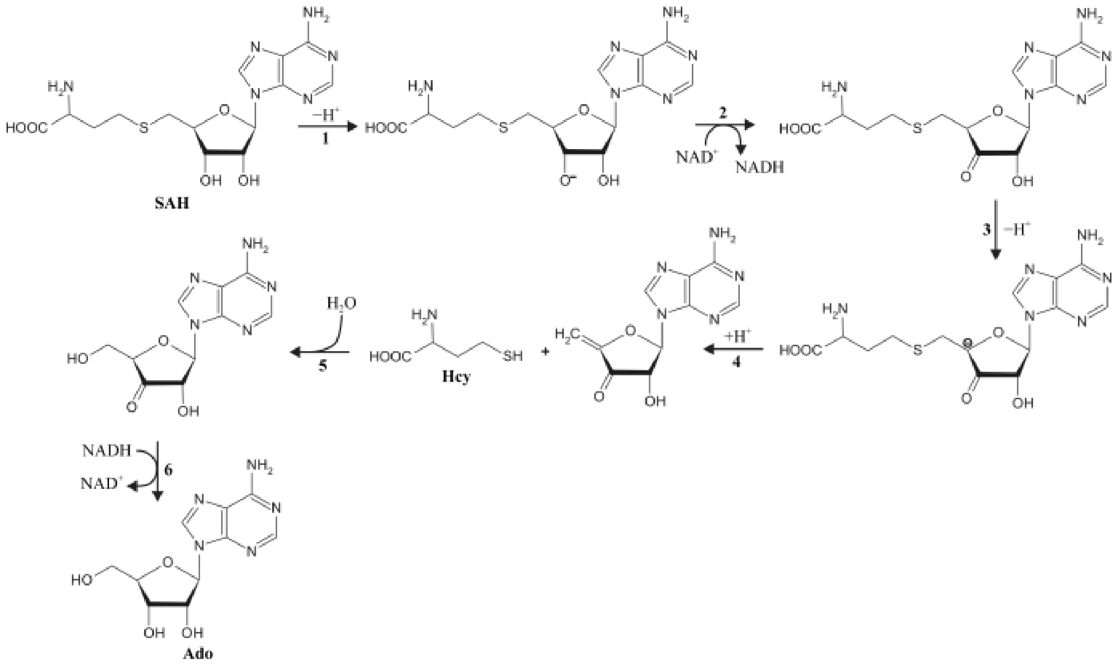


Figure 3.11: Mechanism of SAHH to Ado and Hcy catalysed by S-adenosyl-l-homocysteine hydrolase. The numbers 1–6 indicate the reaction steps described in the mechanism of SAHH. Ado refers to adenosine and Hcy refers to homocysteine. NAD+ is a proton acceptor and NADH is a proton donor. From: Brzezinski *et al.* (2012).

4. Proton transfer from the imidazole ring of histidine62 to the $S\delta$ atom of the 3´-ketoadenosine-homocysteine carbo-anion is followed by β-elimination of homocysteine, leading to the formation of 3´-keto-4´,5´-di-dehydro-adenosine;
5. In the final step, a water molecule is attached through Michael-type addition and,
6. Consequently, the 3´-keto group is reduced and adenosine and NAD^+ are generated (Brzezinski *et al.*, 2012).

Earlier studies have shown that there are differences in kinetic and thermodynamic parameters between the human and parasite enzyme, *Leishmania donovani* SAHH. *L. donovani* SAHH has a weaker binding affinity for NAD^+ and lower catalytic activity, compared to the human enzyme. The *H. sapiens* SAHH tightly binds one NAD^+ cofactor per subunit; this is essential for the catalysis. Due to this differences, inhibitors have been designed to irreversibly reduce the NAD^+ form (active) to the NADH form (inactive) (Khare *et al.*, 2012).

3.4.1.2 Diseases

Many pathological disease states have been related to altered SAH function:

- Hypomethylation of DNA and high homocysteine/SAH levels were shown to be associated with the pathology of cardiovascular diseases in mammals;
- Mice deficient in methylene tetra-hydrofolate reductase, necessary for homocysteine to methionine remethylation, exhibit hyper-homo-cysteinemia and decreased methylation capacity along with neuropathology and aortic lipid deposition and,
- A defect in hepatic phosphatidyl-ethanolamine to phosphatidyl-choline-methylation leads to liver steatosis in mice and was shown to be associated with diabetes in mice and rats (Brzezinski *et al.*, 2012).

Complete loss of SAH function in mammals can also be deleterious for growth and development:

- Deletion of the SAH1 locus in embryonic mice is lethal. Recently patients deficient in Sah1 and exhibiting only 3 – 20% of mean control Sah1 activity were identified that displayed severe myopathy and mental retardation (Brzezinski *et al.*, 2012).

In addition to SAH catabolism, Sah1 plays an important role in homocysteine production, which is required for cysteine and glutathione synthesis. Sah1 produces homocysteine exclusively in mammalian cells (Brzezinski *et al.*, 2012).

3.4.1.3 SAHH in malaria

P. falciparum, *Leishmania donovani* and *Trypanasoma cruzi* have their own SAHH for coping with the toxicity of SAH, so that this target offers opportunities for chemotherapy if structural differences are exploited between the parasite and human enzymes (Khare *et al.*, 2012; Nakanishi, 2007). Nakanishi was busy in 2007 doing research on *P. falciparum*, focusing on the development of new antimalarials based on the SAHH inhibition. *In vitro* screens of nucleoside analogs resulted in moderate but selective inhibition for recombinant SAHH of *P. falciparum*, by 2-position substituted adenosine analogs. Similar selectivity was observed in the growth inhibition assay of cultured cells. Following crystal structure analysis of the parasite SAHH an additional space was discovered, which is located near the 2-position of the adenine-ring, in the substrate-binding pocket. Mutagenic analysis of the amino acid residue forming the additional space confirmed that the inhibition selectivity is due to the difference of only one amino acid residue, between cysteine59 in *P. falciparum* and threonine60 in human. For developing antimalarial drugs, it is ideal to select pathways that are present in the parasite but absent from humans; but even if the target was common in the parasite and the host, slight structural difference, such as single amino acid variations, might improve the inhibitor's selectivity (Nakanishi, 2007).

3.4.2 DISCOVERY

3.4.2.1 Summary

The "Summary" tab provides the user with the PlasmoDB identifier, which for this protein of PFE1050w: S-adenosyl-L-homocysteine hydrolase will be "PFE1050w". The known aliases are as follows:

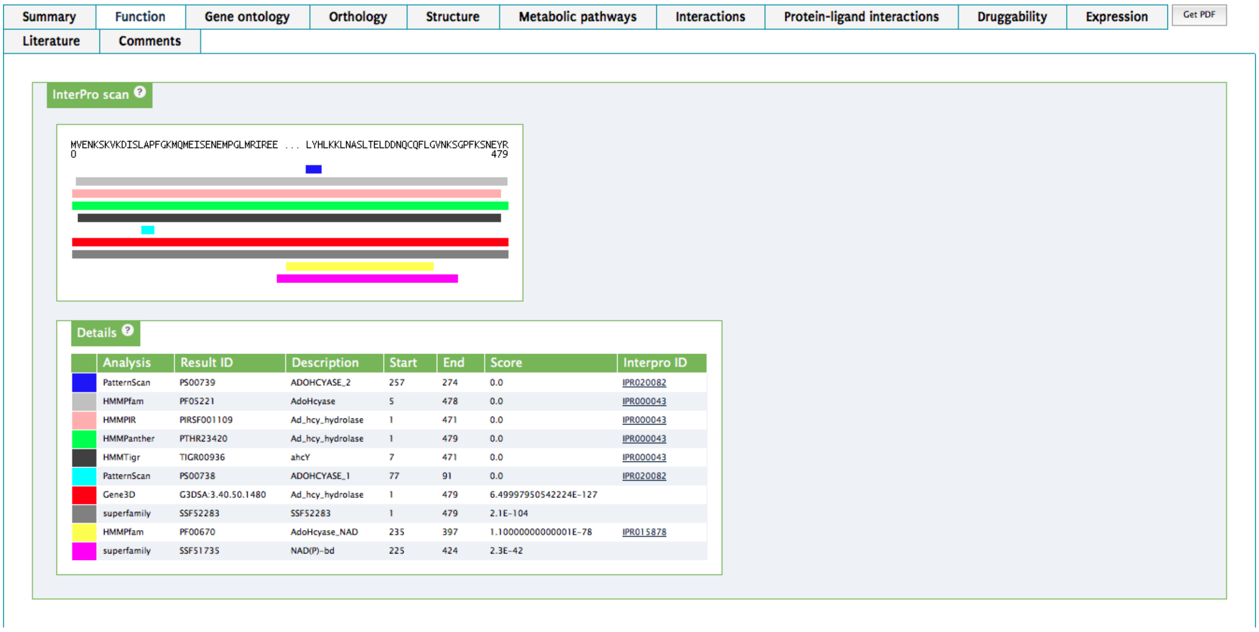
- Adenosylhomocysteinase, S-adenosyl-L-homocysteine hydrolase;
- Adenosylhomocysteinase;
- AdoHcyase;
- PfSAHH and,
- S-adenosyl-L-homocysteine hydrolase.

The Uniprot accession number for this hydrolase is P50250, the number of associated papers is 50. The protein sequence is also given in the "Summary" tab.

3.4.2.2 Function

Figure 3.12 is a screenshot of the "Function" tab. Here it revealed three different Interpro identifiers that matched the query, PFE1050w, they were:

1. IPR020082, S-adenosyl-L-homocysteine hydrolase, conserved site. SAHH is an enzyme of the activated methyl cycle, responsible for the reversible hydration of SAH into adenosine and homocysteine. This enzyme is a highly conserved, omnipresent protein that may play a key role in the



Details

	Analysis	Result ID	Description	Start	End	Score	Interpro ID
🟦	PatternScan	PS00739	ADOHCYASE_2	257	274	0.0	IPR020082
⬜	HMMPfam	PF05221	AdoHcyase	5	478	0.0	IPR000043
🟥	HMMPIR	PIRSF001109	Ad_hcy_hydrolase	1	471	0.0	IPR000043
🟩	HMMPanther	PTHR23420	Ad_hcy_hydrolase	1	479	0.0	IPR000043
⬛	HMMTigr	TIGR00936	ahcY	7	471	0.0	IPR000043
🟦	PatternScan	PS00738	ADOHCYASE_1	77	91	0.0	IPR020082
🟥	Gene3D	G3DSA:3.40.50.1480	Ad_hcy_hydrolase	1	479	6.49997950542224E-127	
⬛	superfamily	SSF52283	SSF52283	1	479	2.1E-104	
🟨	HMMPfam	PF00670	AdoHcyase_NAD	235	397	1.10000000000001E-78	IPR015878
🟪	superfamily	SSF51735	NAD(P)-bd	225	424	2.3E-42	

Figure 3.12: Screenshot of the Function tab for PFE1050w.

Table 3.3: Summary of the InterPro signatures matching to *P. falciparum* SAHH.

InterPro entry	Signatures	Analysis Method
IPR020082 (Conserved site)	PS00739	PatternScabn
	PS00738	PatternScan
IPR000043 (Family)	PF05221	HMMPfam
	PIRSF001109	HMMPIR
	PTHR23420	HMMPanther
	TIGR00936	HMMTigr
IPR015878 (Domain)	PF00670	HMMPfam
Other	G3DSA:3.40.50.1480	Gene3D
	SSF52283	Superfamily
	SSF51735	Superfamily

regulation of the intracellular concentration of adenosyl-homocysteine. SAHH requires NAD^+ as a cofactor and contains a central glycine-rich region which is thought to be involved in the NAD-binding;

2. IPR000043, adenosylhomocysteinase, family. See (1) IPR020082. And,
3. IPR015878, S-adenosyl-L-homocysteine hydrolase, NAD-binding domain. SAHH is a highly conserved protein of about 430 to 470 amino acids.

This entry represents the glycine-rich region in the central part of SAHH, which is thought to be involved in NAD-binding. The InterPro entries can be summarized in Table 3.3.

3.4.2.3 Gene ontology

The "Gene Onthology" tab revealed multiple GO terms associated with the SAHH. They where:

- None at the Cellular component level;
- At the Molecular function level:
  - GO:0004013, Adenosylhomocysteinase activity, with an evidence code of IEA and,
  - GO:0016787, hydrolase activity, with an evidence code of IEA. And,

- At the Biological processes level:
  - GO:0006730, one-carbon metabolic process, with an evidence code of IEA.

3.4.2.4 Orthology

At the "Orthology" tab the OrthoMCl clustering and an Ortholog table can be found. The OrthoMCL clustering showed that the SAHH is present in most of the malaria species in DISCOVERY2 as well as the mosquito and human. In some of the malaria species it is considered as a putative protein, but for most orthologs it is a "real" protein with multiple aliases. All eight species had the same alignment length, but both, the human and mosquito sequences, had a sequence length of 432, 7 gaps and a gap length of 49, one of the gaps can be observed from the T-coffee alignment from database 145 to 187. The *Plasmodium* species had sequence length of 477 or 479, depending on the specie. Those *Plasmodium* species with the sequence length of 477 had a gap length of four, stretched over three gaps while those with a sequence length of 479 had a gap length of two, stretched over two gaps.

3.4.2.5 Structure

The "Structure" tab shows multiple sequences that had a minimum of 70% coverage using PDB BLAST. If one clicks on the sequence link the alignment of the sequence with PFE1050w is shown. There are two possible ModBase structures available for PFE1050w: S-adenosyl-L-homocysteine hydrolase. They are:

- PFE1050w.1, using PDB template 1LI4 and has a DOPE score of -0.92 and,
- PFE1050w.2, using PDB template 1V8B and has a DOPE score of -1.86.

3.4.2.6 Metabolic pathways

In the "Metabolic pathway" tab it is shown that the MPMP pathways associated with PFE1050w are as follows:

- S-glutathionylated proteins with activity S-adenosyl-L-homocysteine adenosylhomocysteinase;
- Methione and polyamine metabolism with activity of adenosylhomocysteinase;
- Total palmitome of *P. falciparum* with activity of adenosyl-homocysteinase (SAHH) and,
- Proteins targeted by the thioredoxin superfamily enzymes with activity of SAHH.

Enzyme information, supplied by the "Metabolic pathway" tab, in the form of EC numbers are as followed for PFE1050w:

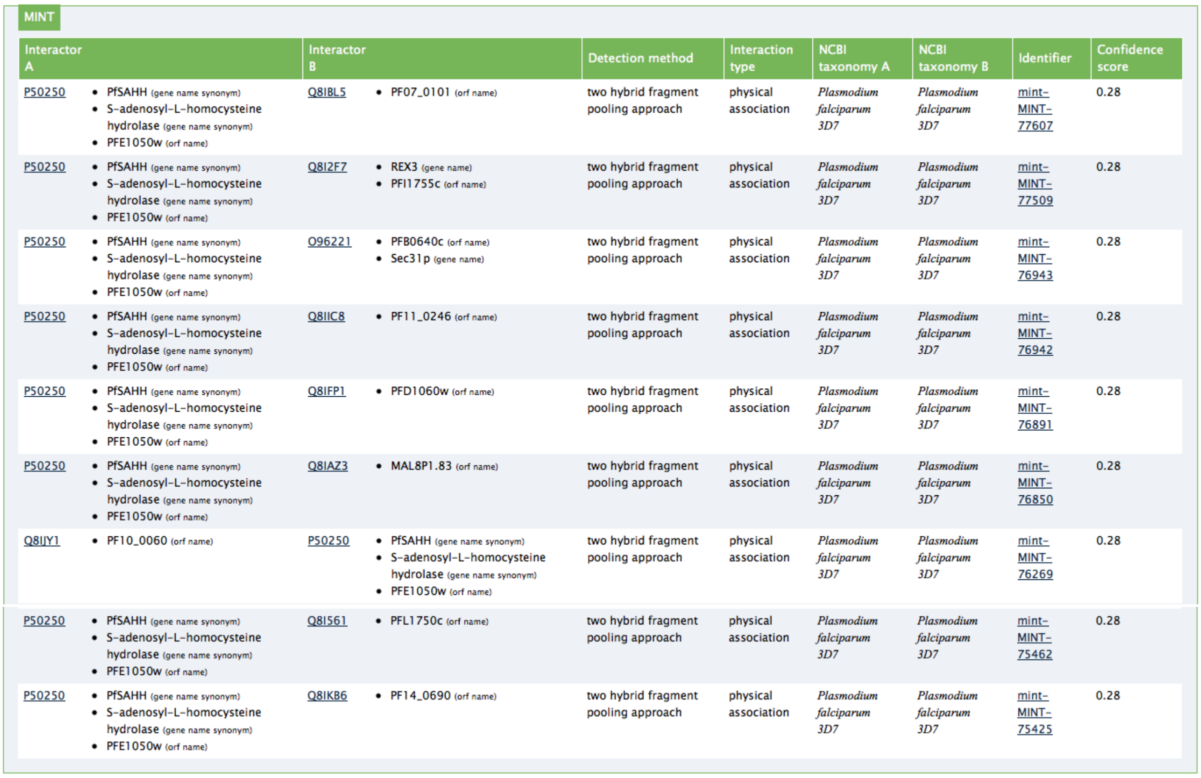
- 3.3.1.1 Adenosylhomocysteinase.

Two KEGG pathways were collected while still publicly available and is:

- Cysteine and methionine metabolism and,
- Metabolic pathways.

Reactome information can also be found under the "Metabolic Pathways" tab, and for PFE1050w it is:

Metabolism
    └Metabolism of amino acids and derivatives
        └Sulfur amino acid metabolism
            └S-adenoylhomocysteine is hydrolyzed (*Plasmodium falciparum*)



MINT

Interactor A	Interactor B	Detection method	Interaction type	NCBI taxonomy A	NCBI taxonomy B	Identifier	Confidence score
P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	Q8IBL5 • PF07_0101 (orf name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-77607	0.28
P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	Q8I2F7 • REX3 (gene name) • PFI1755c (orf name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-77509	0.28
P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	O96221 • PFB0640c (orf name) • Sec31p (gene name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-76943	0.28
P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	Q8IIC8 • PF11_0246 (orf name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-76942	0.28
P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	Q8IFP1 • PFD1060w (orf name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-76891	0.28
P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	Q8IAZ3 • MAL8P1.83 (orf name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-76850	0.28
Q8IJY1 • PF10_0060 (orf name)	P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-76269	0.28
P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	Q8I561 • PFL1750c (orf name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-75462	0.28
P50250 • PfSAHH (gene name synonym) • S-adenosyl-L-homocysteine hydrolase (gene name synonym) • PFE1050w (orf name)	Q8IKB6 • PF14_0690 (orf name)	two hybrid fragment pooling approach	physical association	*Plasmodium falciparum 3D7*	*Plasmodium falciparum 3D7*	mint-MINT-75425	0.28

Figure 3.13: Screenshot of the Interaction tab for PFE1050w.

└Biological oxidations
    └Phase II conjugation
        └Methylation
            └S-adenoylhomocysteine is hydrolyzed (*Plasmodium falciparum*).

3.4.2.7 Interactions

The "Interactions" tab supplies protein-protein interactions and for PFE1050w it has interactions associated with the IntAct and MINT databases (see Figure 3.13 for the MINT interactions). Nine interactions, including it, were associated with PFE1050w, these interacting proteins are:

- UniProt: Q8IBL5, an uncharacterised protein;
- UniProt: Q8I2F7, Ring-exported protein 3 or REX3;
- UniProt: O96221, a putative protein, Sec31p;
- UniProt: Q8IIC8, a conserved *Plasmodium* protein with unknown function;
- UniProt: Q8IFP1, a putative protein, U5 small nuclear ribonucleoprotein-specific protein;
- UniProt: Q8IAZ3, Eukaryotic translation initiation factor 3 subunit G or eIF3g;
- UniProt: Q8IJY1, a conserved protein with unknown function;
- UniProt: Q8I561, a conserved protein with unknown function and,
- UniProt: Q8IKB6, a putative protein, histone deacetylase.

To detect these interactions a two-hybrid fragment pooling approach was used. This technique uses individual pieces, matched against a pool of random fragment. Using degenerated fragments allows identification of the minimal protein region required for an interaction.

3.4.2.8 Protein-ligand interactions

Doing a BLAST search against the ChEMBL targets database identified protein-ligand interactions, as depicted under the “Protein-ligand interactions” tab. Protein-ligand interactions, as depicted under the “Protein-ligand interactions” tab, were identified by doing a BLAST search against the ChEMBL targets database. This is reported in the “By targets” sub tab. In the “By ligands” sub tab the associated ligands are reported.

PFE1050w: S-adenosyl-L-homocysteine hydrolase was identified from the ChEMBL targets database, also included in this result where the SAHH from organisms such as mouse, human and *Thermotoga maritima*. All these organisms had good BLAST results. Domain matching added no additional hits to the full sequence results.

Almost three hundred chemical compounds were matched and have been tested for activity against PFE1050w: S-adenosyl-L-homocysteine hydrolase and its homologs. 819 different bioactivities involving these ~ 300 compounds have been found.

3.4.2.9 Druggability

The “Druggability” tab, as can be seen in Figure 3.14, shows molecules with the most significant BLAST domain matches from DrugEBIlity. For PFE1050w the protein is predicted as druggable. PFE1050w have multiple domains, identified by BLAST that can be considered druggable. There are also multiple genes, identified by BLAST comparisons, with viable scores and E-values to consider PFE1050w: S-adenosyl-L-homocysteine hydrolase as a potential drug target.

3.4.2.10 Expression

In the “Expression” tab it can be seen that PFE1050w: S-adenosyl-L-homocysteine hydrolase is expressed predominantly in the trophozoite stage. It also indicates that PFE1050w is underexpressed in the early ring stage. This may indicate that PFE1050w plays a role in parasite growth and preparation for division, which can also make it a good target.

3.4.2.11 Literature

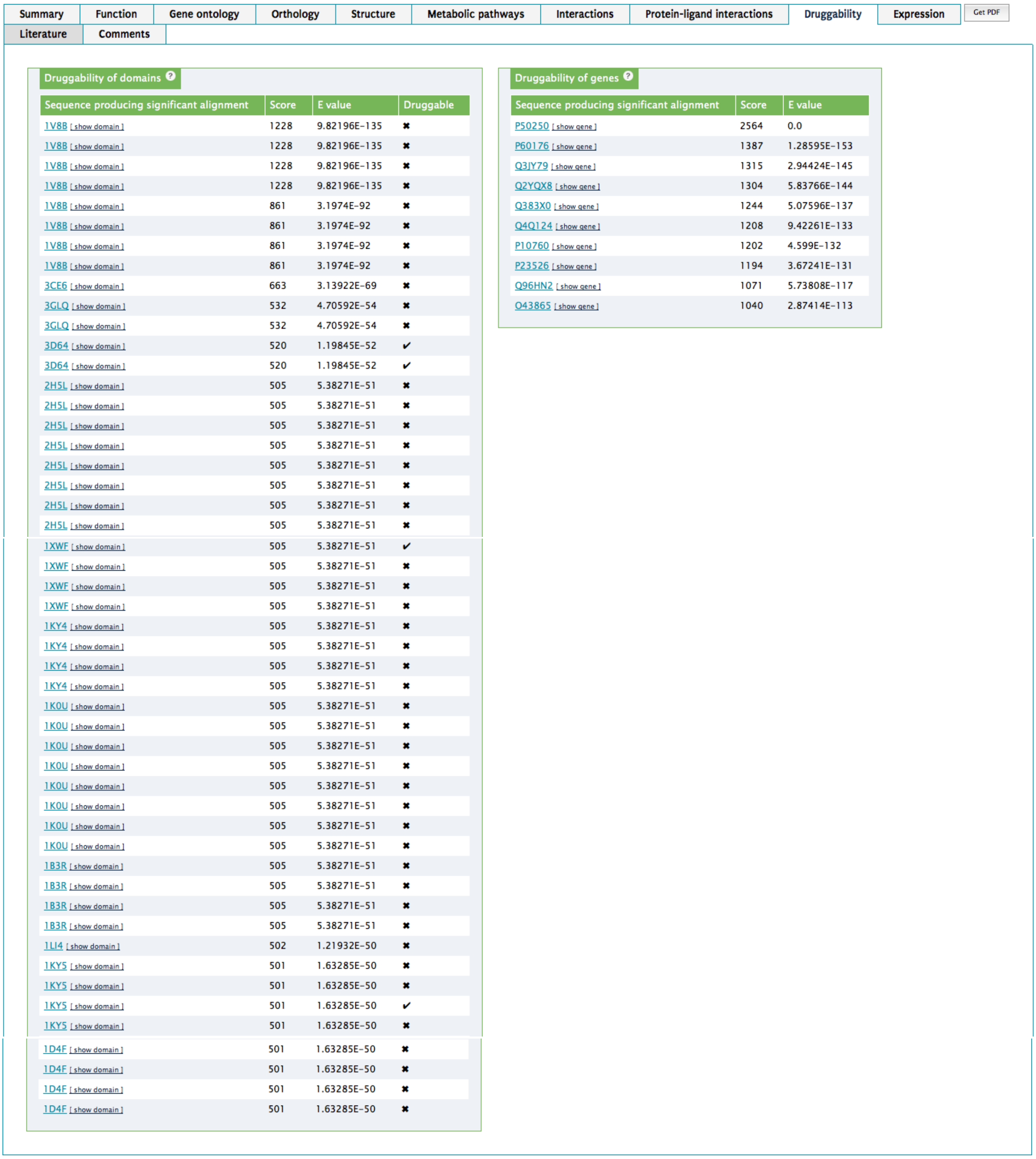
The “Literature” tab showed that there are 50 articles associated with this SAHH. These articles are retrieved from PubMed abstracts, and based on text relation mining of the abstracts it is possible that the following might be interactors of this kinase:

- Aristeromycin;
- Methionine and cysteine;
- Adenine;
- Adenosine, cysteine and ribavirin and,
- Adenosine and imidazole.

Adenosine might have a negative relation with PFE1050w.

3.4.3 Discussion

In DISCOVERY2 there is a lot of data available for PFE1050w: S-adenosyl-L-homocysteine hydrolase; it is a protein that is well conserved across the *Plasmodium* genus, and also across humans and mosquitos, but have differences between the human and parasite form of this protein. Crystal structures are available, possible ligands to bind to this target and possible interactions between this target



Druggability of domains

Sequence producing significant alignment	Score	E value	Druggable
1V8B [show domain]	1228	9.82196E-135	✖
1V8B [show domain]	1228	9.82196E-135	✖
1V8B [show domain]	1228	9.82196E-135	✖
1V8B [show domain]	1228	9.82196E-135	✖
1V8B [show domain]	861	3.1974E-92	✖
1V8B [show domain]	861	3.1974E-92	✖
1V8B [show domain]	861	3.1974E-92	✖
1V8B [show domain]	861	3.1974E-92	✖
3CE6 [show domain]	663	3.13922E-69	✖
3GLQ [show domain]	532	4.70592E-54	✖
3GLQ [show domain]	532	4.70592E-54	✖
3D64 [show domain]	520	1.19845E-52	✔
3D64 [show domain]	520	1.19845E-52	✔
2H5L [show domain]	505	5.38271E-51	✖
2H5L [show domain]	505	5.38271E-51	✖
2H5L [show domain]	505	5.38271E-51	✖
2H5L [show domain]	505	5.38271E-51	✖
2H5L [show domain]	505	5.38271E-51	✖
2H5L [show domain]	505	5.38271E-51	✖
2H5L [show domain]	505	5.38271E-51	✖
2H5L [show domain]	505	5.38271E-51	✖
1XWF [show domain]	505	5.38271E-51	✔
1XWF [show domain]	505	5.38271E-51	✖
1XWF [show domain]	505	5.38271E-51	✖
1XWF [show domain]	505	5.38271E-51	✖
1KY4 [show domain]	505	5.38271E-51	✖
1KY4 [show domain]	505	5.38271E-51	✖
1KY4 [show domain]	505	5.38271E-51	✖
1KY4 [show domain]	505	5.38271E-51	✖
1K0U [show domain]	505	5.38271E-51	✖
1K0U [show domain]	505	5.38271E-51	✖
1K0U [show domain]	505	5.38271E-51	✖
1K0U [show domain]	505	5.38271E-51	✖
1K0U [show domain]	505	5.38271E-51	✖
1K0U [show domain]	505	5.38271E-51	✖
1K0U [show domain]	505	5.38271E-51	✖
1K0U [show domain]	505	5.38271E-51	✖
1B3R [show domain]	505	5.38271E-51	✖
1B3R [show domain]	505	5.38271E-51	✖
1B3R [show domain]	505	5.38271E-51	✖
1B3R [show domain]	505	5.38271E-51	✖
1LI4 [show domain]	502	1.21932E-50	✖
1KY5 [show domain]	501	1.63285E-50	✖
1KY5 [show domain]	501	1.63285E-50	✖
1KY5 [show domain]	501	1.63285E-50	✔
1KY5 [show domain]	501	1.63285E-50	✖
1D4F [show domain]	501	1.63285E-50	✖
1D4F [show domain]	501	1.63285E-50	✖
1D4F [show domain]	501	1.63285E-50	✖
1D4F [show domain]	501	1.63285E-50	✖

Druggability of genes

Sequence producing significant alignment	Score	E value
P50250 [show gene]	2564	0.0
P60176 [show gene]	1387	1.28595E-153
Q3JY79 [show gene]	1315	2.94424E-145
Q2YQX8 [show gene]	1304	5.83766E-144
Q383X0 [show gene]	1244	5.07596E-137
Q4Q124 [show gene]	1208	9.42261E-133
P10760 [show gene]	1202	4.599E-132
P23526 [show gene]	1194	3.67241E-131
Q96HN2 [show gene]	1071	5.73808E-117
O43865 [show gene]	1040	2.87414E-113

Figure 3.14: Screenshot of the Druggability tab for PFE1050w.

and other proteins have been identified, this hydrolase is considered druggable and there have been a lot of research already done on the protein. When PFE1050w was put through Taverna and scored it scored as followed:

Chokepoint:	10
Interaction:	5.2
Ortholog:	2.91
Ligand:	10
ModBase:	10
Literature:	10
Druggability:	10.

Giving it a total weighted score of 0.77, a very promising score that causes this protein to be considered a very good possible drug target for future drug design regarding malaria. If the weights were assigned differently the maximum weighted score PFE1050w would be able to achieve is 0.94 and the minimum would be 0.72. In DISCOVERY PFE1050w had all the available tabs that it can possibly have in DISCOVERY thus data for all criteria where available to get a score. Low scores are less likely to be attributed to no known data. The fact that the protein sequence is very similar to its human ortholog is a negative point.

Researchers have already identified this protein class as a possible antimalarial drug target. Researching the effect ligands might have on this type of protein. This fact identifies the scoring technique as promising Nakanishi (2007).

3.4.4 Conclusion

PFE1050w: S-adenosyl-L-homocysteine hydrolase have a score of 0.77, thus it falls in the 1% that can be considered to have a good score. Researchers have identified this protein as a possible antimalarial drug target, even though it have a low score for being similar to a human ortholog protein, but even small sequence differences can make it a successful drug target against *P. falciparum* with minimal effect against the human host, if these changes are used as binding sites for the drugs.

3.5 PFC0855w: Ubiquitin conjugating enzyme, putative

3.5.1 Introduction

Ubiquitination is a well-characterised post-translational modification that is critical for regulating a wide range of cellular processes in eukaryotic organisms. Sequentially, E1 Ub-activating enzyme, E2 Ub-conjugating enzyme and E3 Ub-ligases carry out the conjugation of Ubiquitin (Ub) to target proteins (Bae and Kim, 2014; Chung *et al.*, 2012). Ubiquitylation normally involves the covalent attachment of an Ub moiety to lysine residues of protein substrates (Chung *et al.*, 2012).

In eukaryotes the endoplasmic reticulum-associated degradation (ERAD) pathway mediates protein degradation during quality control. Deviating proteins are recognised by Endoplasmic Reticulum (ER) luminal chaperone protein disulfide isomerases to help discriminate properly folded from misfolded proteins. Misfolded proteins are shuttled to the DER1 translocon complex, which forms a hydrophobic pore to allow the translocation of the proteins through the ER membrane (see Figure 3.15).

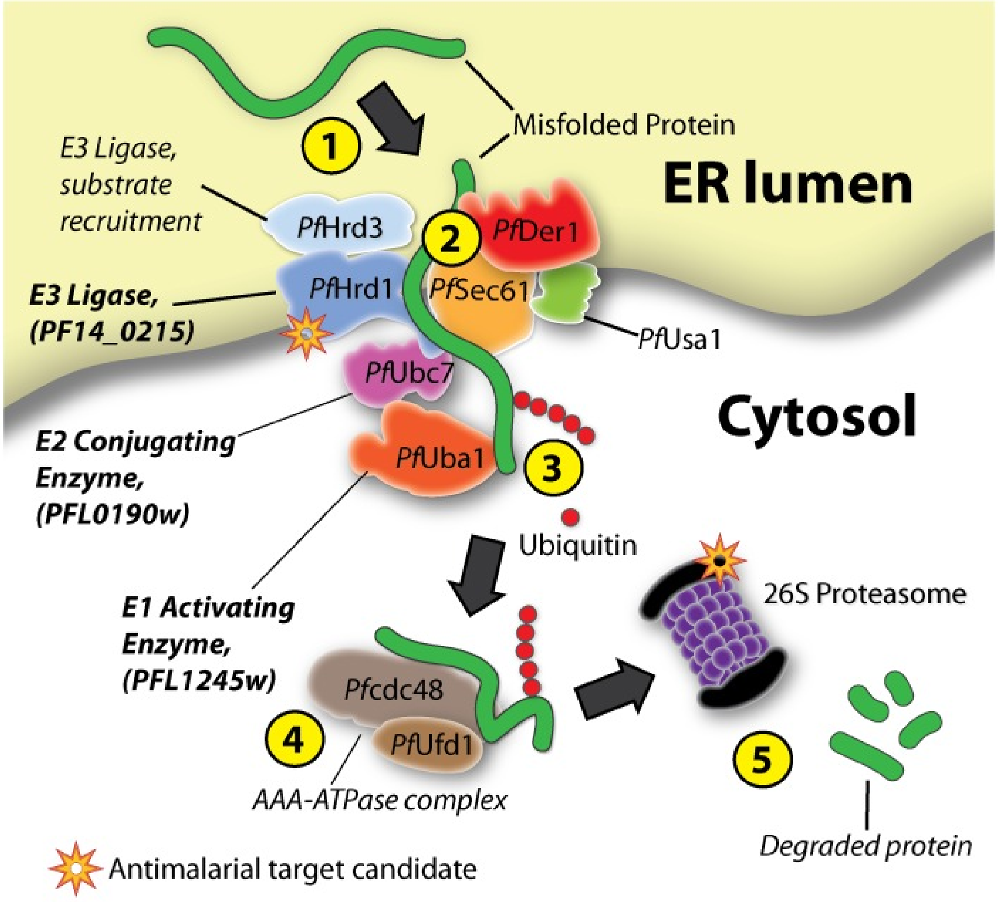


Figure 3.15: Graphical depiction of the Plasmodium ERAD pathway in protein quality control and its potential for antimalarial strategies.
PFL1245w, PFL0190w, and PF14_0215 serve to ubiquitylate misfolded ER proteins for translocation from the ER lumen to the cytosol for proteasome degradation.
From: Chung *et al.* (2012).

3.5.1.1 ERAD in humans

In humans E2s not only interact with E1s and E3s to receive and transfer Ub, respectively, but also regulate, at least in part, the length and topology of the poly-Ub chain and the efficiency of polyubiquitination conducted by E3s. Thus, E2s are involved in ubiquitination as well as are one of the regulators in the ubiquitination pathway (Bae and Kim, 2014).

3.5.1.2 ERAD in malaria

Until recently no functional study has investigated the ubiquitylating components that compose the malaria ERAD pathway. The E1 and E2 Ub enzymes seem to be well conserved across all eukaryotes, while the E3 Ub ligases seem to have high levels of divergences. A recent study of the ERAD pathway in *P. falciparum* suggests that this system is vital to the parasite and that recombinant E1, E2 and E3 enzymes promote ubiquitylation *in vitro*. While immunofluorescence microscopy experiment reveal that E1 and E2 enzymes localizes to the cytosol while E3 is found in the ER membrane, this is consistent with their respective functions in the ERAD pathway. Gene disruption experiments suggest that the ubiquitylating components of the *Plasmodium* ERAD system are essential to the parasite for survival. The component of the ERAD pathway shows promise to be considered as potential targets for antimalarial drugs (Chung *et al.*, 2012).

Chung *et al.* (2012) have shown that Eeyarestatin I, a known ERAD inhibitor, is toxic to the *Plasmodium* parasite within low µM ranges and causes increased levels of ubiquitylated products, indicating that an Ub-dependent ERAD system exists within the parasite, making it a point for antimalarial drug targeting (Chung *et al.*, 2012).

After bioinformatic analysis, *in vitro* ubiquitylation assays and localization studies have been performed it seems that the *Plasmodium* ERAD system may function similarly to that of other eukaryotic model systems. The domain architectures of these proteins in the parasite are homologues to that of there identified counter parts in other eukaryotes and are found in the expected cytosol and ER membrane destinations (Chung *et al.*, 2012).

It may be argued that targeting a highly conserved system, like the ERAD pathway in *P. falciparum*, is likely to have cross-reactivity with similar human host ERAD counterparts and might unlikely produce a tight parasite-specific drug. But, inhibitors of Ub E1 enzymes have shown to reduce leukemia and multiple myeloma, while maintaining limited toxicity in mouse models. The limited toxicity observed in normal host cells might be explained by the notion that inhibition of proteasome degradation has more of a profound effect on rapid growing and multiply cells, such as malignant tumours and protozoan parasites, due to their high metabolic needs (Chung *et al.*, 2012).

Currently, ranges of ubiquitylating enzymes are being screened for inhibitors that may confer anticancer properties. Several inhibitors of Ub and Ub-like enzymes have shown effective results against cancer and are already in use in clinical trials (Chung *et al.*, 2012). For example, studies have revealed a potential role of Ub-conjugating enzyme E2 N (UBE2N) as a novel regulator of p53 by reducing its transcriptional activity. UBE2N associates with p53 and increases the cytoplasm pool of monomeric p53, thereby inhibiting nuclear tetramerization of p53. Due to the critical role of UBE2N in cytosolic sequestration of wild-type p53 and due to the fact that cytosolic sequestration of wild-type p53 is present in most neuroblastoma cancer cases, it is hypothesized that pharmacological inhibition of UBE2N could promote p53 nuclear translocation and its subsequent activation in neuroblastoma (Cheng *et al.*, 2014).

Theoretically, components of the *Plasmodium* ERAD degradation pathway could produce an effective new strategy against *Plasmodium* as they are both specific and likely essential to the parasite life cycle (Chung *et al.*, 2012). Recently a duplicated ERAD-like system has been identified within the plastids of some apicomplexans, cryptomonads and diatoms, including *P. falciparum*. But instead of functioning in protein degradation, these plastid ERAD-like systems is believed to have roles in plastid protein import, which may also turn out to be an effective drug target (Chung *et al.*, 2012).

3.5.2 DISCOVERY

3.5.2.1 Summary

The "Summary" tab as previously seen provides the user with the protein identifier from PlasmoDB, which for PFC0855w: Ubiquitin conjugating enzyme, putative will be "PFC0855w". The alias for this protein is "Ubiquitin conjugating enzyme, putative". Links to Uniprot, PlasmoDB and the KEGG databases are also given. The number of associated articles for PFC0855w is two and as usual the protein sequence for this protein is also supplied.

3.5.2.2 Function

In the "Function" tab two different InterPro identifiers were matched. They were as follows:

1. IPR000608, Ubiquitin-conjugating enzyme, E2. The post-translational attachment of Ub to proteins (ubiquitinylation) alters the function, location or trafficking of a protein, or targets it to the 26S proteasome for degradation. The ubiquitin-activating (E1) enzyme mediates an ATP-dependent transfer of a thioester-linked Ub molecule to a cysteine residue on the ubiquitin-conjugating (E2) enzyme. The E2 enzyme then either transfers the Ub moiety directly to a substrate, or to an Ub (E3) ligase, which can also ubiquitinylate a substrate. There are roughly four classes of E2 enzymes, all of them have a core catalytic domain, and some of which have short N- and C-terminal amino acid extensions:

   (a) Class I enzymes consist of just the catalytic core domain (UBC);
   (b) Class II possess a UBC and a C-terminal extension;
   (c) Class III possess a UBC and an N-terminal extension and,
   (d) Class IV possess a UBC and both N- and C-terminal extensions. These extensions appear to be important for some subfamily function, including E2 localisation and protein-protein interactions. And,

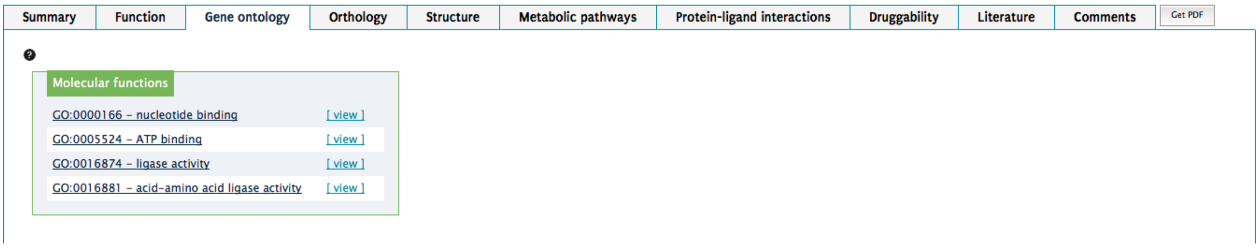
2. IPR016135, Ubiquitin-conjugating enzyme/RWD-like. This identifier represents a structural domain with an alpha-beta(4)-alpha(3) core fold. Domains of this structure can be found in:

   (a) Ubiquitin-conjugating enzyme E2, as well as related proteins such as Ub-carrier protein 4 and ubiquitin-protein ligase W;
   (b) The UEV domain in tumour susceptibility gene 101 and vacuolar protein sorting associated protein;
   (c) RWD domain, found in RING finger and WD repeat-containing proteins, such as EIF-2 kinase 4 (GCN2-like protein) and,
   (d) UFC1-like domain found in Ufm1-conjugating enzyme 1.

The InterPro entries for PFC0855w can be seen in Table 3.4.

Table 3.4: Summary of the InterPro signatures matching to ubiquitin-conjugating enzyme.

InterPro entry	Signatures	Analysis method
IPR000608 (Domain)	PSS0127	ProfileScan
	PS00183	PatternScan
	SM00212	HMMSmart
	PF00179	HMMPfam
IPR06135 (Domain)	SSF54495	superfamily
	G3DSA:3.10.110.10	Gene3D
Other	PTHR11621	HMMPanther
	PTHR11621:SF33	HHMPanther



Molecular functions

GO:0000166 – nucleotide binding [view]
GO:0005524 – ATP binding [view]
GO:0016874 – ligase activity [view]
GO:0016881 – acid-amino acid ligase activity [view]

Figure 3.16: Screenshot of the Gene Onthology tab for PFC0855w.

3.5.2.3 Gene ontology

The "Gene Ontology" tab, can be seen in Figure 3.16. It identified the following GO-terms all at the molecular function level:

- GO:0000166 (nucleotide binding), with an evidence code of IEA;
- GO:0005524 (ATP binding), with an evidence code of IEA;
- GO:0016874 (ligase activity) , with an evidence code of IEAand,
- GO:0016881 (acid-amino acid ligase activity), with an evidence code of IEA.

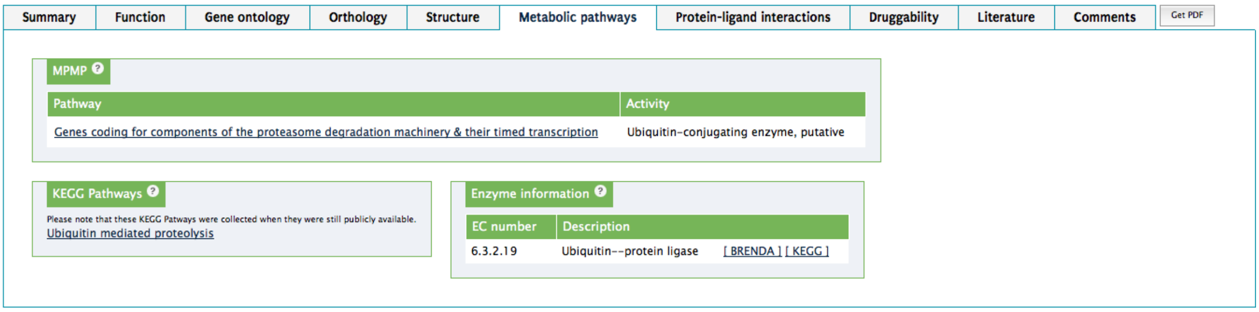
3.5.2.4 Orthology

The "Orthology" tab gives a T-Coffee alignment of all the similar proteins in the different organisms included in DISCOVERY2. From the Ortholog table it was observed that all the sequences had an alignment length of 160 and no gaps where detected. No *H. sapiens* and *A. gambiae* sequences were included in this alignment. Also all other *Plasmodium* spp. was similarly conserved compared to each other.

3.5.2.5 Structure

The "Structure" tab gives a BLAST score table and the ModBase structures. In DISCOVERY PFC0855w is aligned against similar proteins that have known crystal structures and what is shown is the 70% and higher coverage top scoring alignments of this BLAST search. Three viable looking ModBase structures are available for this protein, namely:

- PFC0855w.1, using the 2AAK PDB template, having a DOPE score of −1.98-;
- PFC0855w.2, using the 1YF9 PDB template, having a DOPE score of −1.4 and,
- PFC0855w.3 using the 1YH2 PDB template, having a DOPE score of −2.1.



MPMP

Pathway	Activity
Genes coding for components of the proteasome degradation machinery & their timed transcription	Ubiquitin-conjugating enzyme, putative

KEGG Pathways

Please note that these KEGG Patways were collected when they were still publicly available.
Ubiquitin mediated proteolysis

Enzyme information

EC number	Description
6.3.2.19	Ubiquitin--protein ligase [BRENDA] [KEGG]

Figure 3.17: Screenshot of the Metabolic pathways tab for PFC0855w.

3.5.2.6 Metabolic pathways

The Metabolic pathways tab gives the MPMP pathways, KEGG pathways and Enzyme information as can be seen in Figure 3.17. The MPMP database shows that it is associated with genes coding for components of the proteasome degradation machinery and it is timed transcription with activity of ubiquitin-conjugating enzyme. The KEGG pathway associates PFC0855w with ubiquitin-mediated proteolysis (*these pathway were collected when KEGG was still publicly available). And the enzyme information give the EC number as 6.3.2.19, an ubiquitin-protein ligase.

3.5.2.7 Protein-ligand interactions

The "Protein-ligand interactions" tab shows the Protein-ligand interaction as it was identified by a BLAST search against the ChEMBL targets database and reported in the target section. The associated ligands are reported in the ligand section. From the targets database two sequences produced significant alignments and six sequences produced domain matches. All six sequences had the same two domain matches, namely: UBQconjugate/ RWD-like and UQ_con. One domain-matched sequence had no protein associations. From the ligand database more than 1,000 chemical compounds were matched that have been tested against PFC0855w and its homologs.

3.5.2.8 Druggability

Druggability data from the "Druggability" tab showed that the molecules with the most significant BLAST domain matches were not predicted to be druggable by DrugEBIlity, although there are some domains with a significant BLAST score predicted to be druggable, thus PFC0855w: Ubiquitin conjugating enzyme, putative can be considered as a druggable protein.

3.5.2.9 Literature

The "Literature" tab showed two articles extracted from PubMed abstracts related to the protein. Text relation mining of abstracts showed that the following compounds might interact with PFC0855w:

- Aspermine;
- Lysine;
- Substrate A;
- Adenosine;
- Progesterone and,
- Cisplatin.

Of these interactors lysine can interact with it in a negative way.

3.5.3 Discussion

In DISCOVERY2 there is a quite a bit of data available for PFC0855w: Ubiquitin conjugating enzyme, putative; it is a protein that is very well conserved across the *Plasmodium* genus, and has again no orthologs in humans and mosquitos; crystal structures are available; there are possible ligands to bind to this target; it is considered druggable and literature data shows that there might be a product that interacts negatively with this *Plasmodium* protein. When PFC0855w was put through Taverna and scored it scored as followed:

Chokepoint:	10
Interaction:	10
Ortholog:	10
Ligand:	10
ModBase:	10
Literature:	5
Druggability:	10.

This gives it a total weighted score of 0.96. If the weights were assigned differently the maximum weighted score PFC0855w would be able to achieve is 0.98 and the minimum would be 0.88. Thus PFC0855w: Ubiquitin-conjugating enzyme, putative can be considered an excellent possible drug target for future drug design regarding malaria. Researchers have already started to look into the properties of this type of protein as a possible antimalarial drug target (Chung *et al.*, 2012).

PFC0855w was chosen for a case study as it had received the highest score of all the *P. falciparum* proteins. The amount of data available on DISCOVERY was less than the two intermittent proteins, and only had protein-ligand interactions extra, compared to PF11_0058-b, the case study protein with a score of 0.16. But all the data for the observed criteria for PFC0855w was positive, compared to PF11_0058-b where the data was sometimes less than ideal.

The ERAD system, of which PFC0855w is part of, has already been identified as a potential drug target by researchers and functional studies are being researched. This, again, identifies the scoring technique as promising (Chung *et al.*, 2012).

3.5.4 Conclusion

PFC0855w: Ubiquitin-conjugating enzyme, putative have a score of 0.96. It scored the best of all the 5,000 plus protein available in *P. falciparum* and because the score is bigger than 0.77 PFC0855w can be considered to have a good score. Researchers have already identified the system of which this protein is part of as possible future drug targets against *P. falciparum*, but more research might be needed before possible drugs against these proteins can enter clinical trials.

3.6 Discussion

Ideally the in-house scoring technique would have a higher score for the proteins that are more likely to be considered future drug targets for antimalarial drugs. This was observed in these four case studies. Of the four proteins mentioned the two proteins with the highest score, PFC0855w: Ubiquitin conjugating enzyme, putative and PFE1050w: S-adenosyl-L-homocysteine hydrolase seemed the most likely proteins to be included in the next generation antimalarial drug targets as they had scores that can be considered good scores. But these two proteins are part of protein types already involved in research regarding new antimalarial drug targets, this is encouraging for the criteria identified and the scores allocated to each.

Due to little information being available for PF11_0058-b: RNA polymerase subunit, putative it had a low score. And for this reason it should ideally not be considered a possible antimalarial drug target, until more information become available, and even then it might not score highly. Even though more information is available regarding PFL2280w: serine/threonine PK, putative, not all of the criteria had a favourable score, making it not an ideal drug target according to the scoring technique. The point that influenced this protein the most negatively is the fact that it is not considered a chokepoint, even though it can be considered as a druggable protein. More information regarding this protein might change the score.

3.7 Conclusion

The in-house scoring technique functioned competitively. It scored known antimalarial drug targets with a score bigger than 0.77, a score that is considered a good score. The two proteins studied in the case study with scores bigger than 0.77, PFC0855w: Ubiquitin conjugating enzyme, putative and PFE1050w: S-adenosyl-L-homocysteine hydrolase, is part of protein types that are considered for future drug targets, strengthening the in-house scoring system. Little information is known about PF11_0058-b: RNA polymerase subunit, putative thus it had a low score it is also not considered a future drug target, and would doubtful ever be due to its function being needed in all organisms, and even though there is a lot of information available for PFL2280w: serine/threonine PK it scored mediocre due to one single scoring point. PFL2280w has not yet been identified as a possible as a future drug target, this might be due to it not considered a chokepoint, further strengthening the in-house scoring system.

Chapter 4

Concluding Discussion

Most antimalarial drugs are showing a degree of resistance, due to the over-and misuse of antimalarial drugs thus new antimalarial drugs need to be discovered. But to manually sift through the 200 plus possible *Plasmodium* proteins that have the possibility of being drug targets can be quite a lengthy process. The DISCOVERY Database aims to aid in this filtering process by being a database filled with aspects that might influence the druggability of a protein and guide a scientist in choosing the right ligand for that protein.

The main aim of this project was to identify proteins of *P. falciparum* that have a possibility of being druggable. This was achieved by running all the *P. falciparum* proteins through the DISCOVERY database; by wrapping the DISCOVERY web services and using it in a workflow pipeline, constructed in Taverna; and scoring the proteins with a weighted in-house scoring technique. An aggregate score was composed and the proteins were ranked according to this aggregate score.

This study is aimed to contribute towards drug target discovery in the malaria parasite by having a ranked list of proteins that might have a possibility of being the future drug targets. Three proteins, which are known to be drug targets were used as positive controls in the study. These proteins were:

- PF14_0641, 1-deoxy-D-xylulose-5-phosphate reductoisomerase. *In vitro* studies have suggested fosmidomycin as a drug that can target PF14_0641, and be effective against *P. falciparum* (Jomaa *et al.*, 1999; Yeh *et al.*, 2004);
- PFD0830w, Dihydrofolate reductase is known to be essential for the malaria parasite and are a a target for pyrimethamine and cycloguanil, but mutation in this gene has caused some resistance (Yeh *et al.*, 2004; Yuthavong *et al.*, 2012). And,
- PF08_0095, Dihydropteroate synthase is a target for sulfone/sulfonamide drugs, but mutations have been observed for at least the past 10 years (Triglia *et al.*, 1997; Yeh *et al.*, 2004).

Scores for all $\sim 5,500$ proteins ranged from 0.01 to 0.96. These three proteins had Final aggregate scores of 0.84, 0.92 and 0.79 respectively.

Scores equal or bigger than 0.77 can be considered a good score as 99% of all scores had a value lower than this, and only 58 proteins (1%) had values of 0.77 or higher. Two of the proteins that were used for case studies, namely PFE1050w: S-adenosyl-L-homocysteine hydrolase (section 3.4) and PFC0855w: Ubiquitin conjugating enzyme, putative (section 3.5) are part of protein classes that have already entered research as potential antimalarial drug targets. These two proteins had scores of 0.77 and 0.96 respectively using the in-house scoring technique. More proteins with significant scores might also already be included in research for new antimalarial development. Indicating that this study can be significant in aiding drug target discovery by supplying a prioritized list of possible drug targets, making the filtering process easier and less labour intensive.

This study is limited by the amount of information available for each of the proteins. This lack of information causes an uncertainty whether the protein received its low score due to lack of information, in effect scoring it as a "false negative", or due to sufficient information, but not ideal data, scoring it as a "true negative". If more information is obtained due to know protein characteristics and available date and less information is predicted, by comparing the protein to known characteristics of other proteins, the weighting and scoring system can be more reliant by weighting the score of, example "Druggability", higher. Also negative scores can be given for unwanted criteria, such as "Human orthologs".

The results from this study can aid in future research by supplying a ranked list of probable new antimalarial drug targets. Filtering through all the potential drug targets can be a lengthy procedure. By using this ranked list the filtering procedure can be minimalized and more time and resources can be spent on optimizing ligands for the potential drug target.

In this study a list of all the known proteins of *P. falciparum* was obtained and weighted and scored, depending on criteria considered important for drug targets. The scoring system produced a ranked list of *P. falciparum* proteins that showed promise in the identification of next generation antimalarial drug targets due to the inclusion of three known antimalarial drug targets and the fact that some of the proteins that had a good score are part of protein types that are already involved in research to identify new drug targets. This list was obtained by creating a workflow system that identified important drug target criteria for the proteins via DISCOVERY. All the known *P. falciparum* proteins was run through the DISCOVERY system, via a Taverna workflow, and scored. Weighing and adding of all the scores for a particular protein was done with the aid of a Python script to compose an Final aggregate score. Proteins with a high score was identified as druggable, compared to proteins with a low score. Four case studies aided as examples as to how good the workflow and scoring system worked.

This study will thus aid researchers by shortening the filtering time by aiding and potentially eliminating the manual sifting procedure to identify potential drug targets. It can also aid all the people in malaria endemic areas by providing the starting off point for the identification of the next generation antimalarial drug targets.

Summary

A drug target has been defined by Imming *et al.* (2006) as "*a target to be a molecular structure (chemically definable by at least a molecular mass) that will undergo a specific interaction with chemicals that we call drugs because they are administered to treat or diagnose a disease. The interaction has a connection with the clinical effect(s).*" Due to the overall misuse of drugs drug resistance has made an appearance everywhere, leading to an increased urge to identify new drug targets. The raise in *in silico* methods aided in this identification, by decreasing the time used to manually filter through proteins in a laboratory.

Drug resistance in malaria has been observed for virtually all drugs available, and if there is no resistance yet for that drug it is rapidly increasing. Thus to identify new antimalarial drug targets is of importance. This study aids in doing just that by constructing a ranked list of potential antimalarial drug targets. This was achieved by running all *P. falciparum* proteins through DISCOVERY in a high throughput manner using a Taverna pipeline. A Python program was then used to score protein criteria that are considered important for drug targets. These criteria included chokepoints, known interactions, human orthologs, known ligands, ModBase structures, available literature and druggability. The closer the protein had a score to one, the more likely it is to be a potential drug target against malaria.

A score of 0.77 and bigger was considered a good score as only 1% of all the aggregate scores had values equal or bigger than 0.77, or 99% of all values was smaller. This came to 58 proteins, thus 58 proteins can be considered a good potential drug target. These 58 proteins included the three proteins that are known antimalarial drug targets and also proteins that are already being researched for future antimalarial drug targets, indicating that the in-house scoring system designed for this study functioned competitively.

Bibliography

Fernan Aguero, Bissan Al-Lazikani, Martin Aslett, Matthew Berriman, Frederick S Buckner, Robert K Campbell, Santiago Carmona, Ian M Carruthers, A W Edith Chan, Feng Chen, Gregory J Crowther, Maria A Doyle, Christiane Hertz-Fowler, Andrew L Hopkins, Gregg McAllister, Solomon Nwaka, John P Overington, Arnab Pain, Gaia V Paolini, Ursula Pieper, Stuart A Ralph, Aaron Riechers, David S Roos, Andrej Sali, Dhanasekaran Shanmugam, Takashi Suzuki, Wesley C Van Voorhis, and Christophe L M J Verlinde. Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat Rev Drug Discov*, 7(11):900–7, Nov 2008.

Daniel Alvarez-Garcia, Jesus Seco, Peter Schmidtke, and Xavier Barril. *Protein-Ligand Interactions*, chapter 13. Druggability Prediction, pages 265–282. Wiley-VCH, 1 edition, 2012.

Marcelo Alves-Ferreira, Ana Carolina Ramos Guimaraes, Priscila Vanessa da Silva Zabala Capriles, Laurent E Dardenne, and Wim M Degrave. A new approach for potential drug target discovery through *in silico* metabolic pathway analysis using *Trypanosoma cruzi* genome information. *Mem Inst Oswaldo Cruz*, 104(8):1100–10, Dec 2009.

Polamarasetty Aparoy, Kakularam Kumar Reddy, and Pallu Reddanna. Structure and ligand based drug design strategies in the development of novel 5- LOX inhibitors. *Curr Med Chem*, 19(22): 3763–78, 2012.

Cristina Aurrecoechea, John Brestelli, Brian P Brunk, Jennifer Dommer, Steve Fischer, Bindu Gajria, Xin Gao, Alan Gingle, Greg Grant, Omar S Harb, Mark Heiges, Frank Innamorato, John Iodice, Jessica C Kissinger, Eileen Kraemer, Wei Li, John A Miller, Vishal Nayak, Cary Pennington, Deborah F Pinney, David S Roos, Chris Ross, Christian J Stoeckert, Jr, Charles Treatman, and Haiming Wang. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res*, 37(Database issue):D539–43, Jan 2009.

Hansol Bae and Woo Taek Kim. Classification and interaction modes of 40 rice E2 ubiquitin-conjugating enzymes with 17 rice ARM-U-box E3 ubiquitin ligases. *Biochem Biophys Res Commun*, 444(4):575–80, Feb 2014.

Tala M Bakheet and Andrew J Doig. Properties and identification of human protein drug targets. *Bioinformatics*, 25(4):451–7, Feb 2009.

Vangie Beal. Web services, N/A. URL `http://www.webopedia.com/TERM/W/Web_Services.html`.

Oren M Becker, Yael Marantz, Sharon Shacham, Boaz Inbal, Alexander Heifetz, Ori Kalid, Shay Bar-Haim, Dora Warshaviak, Merav Fichman, and Silvia Noiman. G protein-coupled receptors: *in silico* drug discovery in 3D. *Proc Natl Acad Sci U S A*, 101(31):11304–9, Aug 2004.

Pramod C Bhasme, Mahantesh M Kurjogi, Rajeshwari D Sanakal, Rohit B Kaliwal, and Basappa B Kaliwal. *In silico* characterization of putative drug targets in *Staphylococcus saprophyticus*, causing bovine mastitis. *Bioinformation*, 9(7):339–44, 2013.

Lyn-Marie Birkholtz, Olivier Bastien, Gordon Wells, Delphine Grando, Fourie Joubert, Vinod Kasam, Marc Zimmermann, Philippe Ortet, Nicolas Jacq, Nadia Saidani, Sylvaine Roy, Martin Hofmann-Apitius, Vincent Breton, Abraham I Louw, and Eric Marechal. Integration and mining of malaria molecular, functional and pharmacological data: how far are we from a chemogenomic knowledge space? *Malar J*, 5:110, 2006.

Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–403, Sep 2009.

Krzysztof Brzezinski, Zbigniew Dauter, and Mariusz Jaskolski. High-resolution structures of complexes of plant S-adenosyl-L-homocysteine hydrolase (*Lupinus luteus*). *Acta Crystallogr D Biol Crystallogr*, 68(Pt 3):218–31, Mar 2012.

Dong-Sheng Cao, Shao Liu, Qing-Song Xu, Hong-Mei Lu, Jian-Hua Huang, Qian-Nan Hu, and Yi-Zeng Liang. Large-scale prediction of drug-target interactions using protein sequences and drug topological structures. *Anal Chim Acta*, 752:1–10, Nov 2012.

Jane M Carlton, John H Adams, Joana C Silva, Shelby L Bidwell, Hernan Lorenzi, Elisabet Caler, Jonathan Crabtree, Samuel V Angiuoli, Emilio F Merino, Paolo Amedeo, Qin Cheng, Richard M R Coulson, Brendan S Crabb, Hernando A Del Portillo, Kobby Essien, Tamara V Feldblyum, Carmen Fernandez-Becerra, Paul R Gilson, Amy H Gueye, Xiang Guo, Simon Kang'a, Taco W A Kooij, Michael Korsinczky, Esmeralda V-S Meyer, Vish Nene, Ian Paulsen, Owen White, Stuart A Ralph, Qinghu Ren, Tobias J Sargeant, Steven L Salzberg, Christian J Stoeckert, Steven A Sullivan, Marcio M Yamamoto, Stephen L Hoffman, Jennifer R Wortman, Malcolm J Gardner, Mary R Galinski, John W Barnwell, and Claire M Fraser-Liggett. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, 455(7214):757–63, Oct 2008.

Sree Krishna Chanumolu, Chittaranjan Rout, and Rajinder S Chauhan. UniDrug-target: a computational tool to identify unique drug targets in pathogenic bacteria. *PLoS One*, 7(3):e32833, 2012.

J Cheng, Y-H Fan, X Xu, H Zhang, J Dou, Y Tang, X Zhong, Y Rojas, Y Yu, Y Zhao, S A Vasudevan, H Zhang, J G Nuchtern, E S Kim, X Chen, F Lu, and J Yang. A small-molecule inhibitor of UBE2N induces neuroblastoma cell death via activation of p53 and JNK pathways. *Cell Death Dis*, 5:e1079, 2014.

Duk-Won D Chung, Nadia Ponts, Jacques Prudhomme, Elisandra M Rodrigues, and Karine G Le Roch. Characterization of the ubiquitylating components of the human malaria parasite's protein degradation pathway. *PLoS One*, 7(8):e43477, 2012.

Geoffrey Cooper. *The Cell: A Molecular Approach*, chapter 6. RNA Synthesis and Processing. Sinauer Associates Inc, 2nd edition, 2000.

Gregory J Crowther, Dhanasekaran Shanmugam, Santiago J Carmona, Maria A Doyle, Christiane Hertz-Fowler, Matthew Berriman, Solomon Nwaka, Stuart A Ralph, David S Roos, Wesley C Van Voorhis, and Fernan Aguero. Identification of attractive drug targets in neglected-disease pathogens using an *in silico* approach. *PLoS Negl Trop Dis*, 4(8):e804, 2010.

T A P de Beer, G A Wells, P B Burger, F Joubert, E Marechal, L Birkholtz, and A I Louw. Antimalarial drug discovery: *in silico* structural biology and rational drug design. *Infect Disord Drug Targets*, 9 (3):304–18, Jun 2009.

Neekesh V Dharia, Arnab Chatterjee, and Elizabeth A Winzeler. Genomics and systems biology in malaria drug discovery. *Curr Opin Investig Drugs*, 11(2):131–8, Feb 2010.

Pierre M Durand, Kubendran Naidoo, and Theresa L Coetzer. Evolutionary patterning: a novel approach to the identification of potential drug target sites in *Plasmodium falciparum*. *PLoS One*, 3 (11):e3685, 2008.

Fredrik N B Edfeldt, Rutger H A Folmer, and Alexander L Breeze. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discov Today*, 16(7-8):284–7, Apr 2011. doi: 10.1016/j.drudis.2011.02.002.

Segun Fatumo, Kitiporn Plaimas, Ezekiel Adebiyi, and Rainer Konig. Comparing metabolic network models based on genomic and automatically inferred enzyme information from *Plasmodium* and its human host to define drug targets *in silico*. *Infect Genet Evol*, 11(1):201–8, Jan 2011.

Mark W E J Fiers, Ate van der Burgt, Erwin Datema, Joost C W de Groot, and Roeland C H J van Ham. High-throughput bioinformatics with the Cyrille2 pipeline system. *BMC Bioinformatics*, 9: 96, 2008.

Andres F Florez, Daeui Park, Jong Bhak, Byoung-Chul Kim, Allan Kuchinsky, John H Morris, Jairo Espinosa, and Carlos Muskus. Protein network prediction and topological analysis in *Leishmania major* as a tool for drug target selection. *BMC Bioinformatics*, 11:484, 2010.

Zhenting Gao, Honglin Li, Hailei Zhang, Xiaofeng Liu, Ling Kang, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Xicheng Wang, and Hualiang Jiang. PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics*, 9:104, 2008.

Donald L Gardiner, Tina S Skinner-Adams, Christopher L Brown, Katherine T Andrews, Colin M Stack, James S McCarthy, John P Dalton, and Katharine R Trenholme. *Plasmodium falciparum*: new molecular targets with potential for antimalarial drug development. *Expert Rev Anti Infect Ther*, 7(9):1087–98, Nov 2009.

Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W Mather, Akhil B Vaidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G Mani Subramanian, Chris Mungall, J Craig Venter, Daniel J Carucci, Stephen L Hoffman, Chris Newbold, Ronald W Davis, Claire M Fraser, and Bart Barrell. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511, Oct 2002.

Dharmendra B Goswami, Lisa M Ogawa, Joshua M Ward, Gregory M Miller, and Eric J Vallender. Large-scale polymorphism discovery in macaque G-protein coupled receptors. *BMC Genomics*, 14: 703, 2013.

Neil Hall, Marianna Karras, J Dale Raine, Jane M Carlton, Taco W A Kooij, Matthew Berriman, Laurence Florens, Christoph S Janssen, Arnab Pain, Georges K Christophides, Keith James, Kim Rutherford, Barbara Harris, David Harris, Carol Churcher, Michael A Quail, Doug Ormond, Jon Doggett, Holly E Trueman, Jacqui Mendoza, Shelby L Bidwell, Marie-Adele Rajandream, Daniel J Carucci, John R Yates, 3rd, Fotis C Kafatos, Chris J Janse, Bart Barrell, C Michael R Turner, Andrew P Waters, and Robert E Sinden. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, 307(5706):82–6, Jan 2005.

Samiul Hasan, Sabine Daugelat, P S Srinivasa Rao, and Mark Schreiber. Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. *PLoS Comput Biol*, 2(6):e61, Jun 2006.

Brian S Hilbush, John H Morrison, Warren G Young, J Gregor Sutcliffe, and Floyd E Bloom. New prospects and strategies for drug target discovery in neurodegenerative disorders. *NeuroRx*, 2(4): 627–37, Oct 2005.

Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34(Web Server issue):W729–32, Jul 2006.

IBM. Create a Web service wrapper, 2013. URL `http://pic.dhe.ibm.com/infocenter/tpfhelp/current/index.jsp?topic=%2Fcom.ibm.ztpf-ztpfdf.doc_put.cur%2Fgtps6%2Fs6tcre8wservwrapper.html`.

Murat Iskar, Monica Campillos, Michael Kuhn, Lars Juhl Jensen, Vera van Noort, and Peer Bork. Drug-induced regulation of target expression. *PLoS Comput Biol*, 6(9), 2010.

H Jomaa, J Wiesner, S Sanderbrand, B Altincicek, C Weidemeyer, M Hintz, I Türbachova, M Eberl, J Zeidler, H K Lichtenthaler, D Soldati, and E Beck. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science*, 285(5433):1573–6, Sep 1999.

William L Jorgensen. *Drug Design: Structure- and Ligand-Based Approaches*, chapter Preface. Cambridge University Press, 2010.

Fourie Joubert, Claudia M Harrison, Riaan J Koegelenberg, Christiaan J Odendaal, and Tjaart A P de Beer. Discovery: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. *Malar J*, 8:178, 2009.

Konrad J. Karczewski, Roxana Daneshjou, and Russ B. Altman. *Translational Bioinformatics*, volume 8(12), chapter 7: Pharmacogenomics. PLOS Computational Biology, 2012.

Prashant Khare, Amit K Gupta, Praveen K Gajula, Krishna Y Sunkari, Anil K Jaiswal, Sanchita Das, Preeti Bajpai, Tushar K Chakraborty, Anuradha Dube, and Anil K Saxena. Identification of novel S-adenosyl-L-homocysteine hydrolase inhibitors through homology-model-based virtual screening, synthesis, and biological evaluation. *J Chem Inf Model*, 52(3):777–91, Mar 2012.

Baharak Khoshkholgh-Sima, Soroush Sardari, Jalal Izadi Mobarakeh, and Ramezan Ali Khavari-Nejad. An *in Silico* Approach for Prioritizing Drug Targets in Metabolic Pathway of *Mycobacterium tuberculosis*. *World of Academy of Science, Engineering and Technology*, 5(11):2114 – 2117, 2011.

Yuji Koseki, Tomohiro Kinjo, Maiko Kobayashi, and Shunsuke Aoki. Identification of novel antimycobacterial chemical agents through the *in silico* multi-conformational structure-based drug screening of a large-scale chemical library. *Eur J Med Chem*, 60:333–9, Feb 2013.

Honglin Li, Zhenting Gao, Ling Kang, Hailei Zhang, Kun Yang, Kunqian Yu, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Jianhua Shen, Xicheng Wang, and Hualiang Jiang. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res*, 34(Web Server issue):W219–24, Jul 2006.

Qingliang Li and Luhua Lai. Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics*, 8:353, 2007.

W B Li, D J Bzik, H M Gu, M Tanaka, B A Fox, and J Inselburg. An enlarged largest subunit of *Plasmodium falciparum* RNA polymerase ii defines conserved and variable RNA polymerase domains. *Nucleic Acids Res*, 17(23):9621–36, Dec 1989.

Zhenping Li, Rui-Sheng Wang, and Xiang-Sun Zhang. Two-stage flux balance analysis of metabolic networks for drug target identification. *BMC Syst Biol*, 5 Suppl 1:S11, 2011.

Xiaofeng Liu, Sisheng Ouyang, Biao Yu, Yabo Liu, Kai Huang, Jiayu Gong, Siyuan Zheng, Zhihua Li, Honglin Li, and Hualiang Jiang. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res*, 38(Web Server issue): W609–14, Jul 2010.

Dinora Lopes, Kanchana Rungsihirunrat, Fatima Nogueira, Aree Seugorn, Jose Pedro Gil, Virgilio E do Rosario, and Pedro Cravo. Molecular characterisation of drug-resistant *Plasmodium falciparum* from Thailand. *Malar J*, 1:12, Oct 2002.

Avi Ma'ayan and John C He. Protein kinase target discovery from genome-wide messenger RNA expression profiling. *Mt Sinai J Med*, 77(4):345–9, 2010.

Avi Ma'ayan, Sherry L Jenkins, Joseph Goldfarb, and Ravi Iyengar. Network analysis of FDA approved drugs and their targets. *Mt Sinai J Med*, 74(1):27–32, Apr 2007.

Nermina Malanovic, Ingo Streith, Heimo Wolinski, Gerald Rechberger, Sepp D Kohlwein, and Oksana Tehlivets. S-adenosyl-l-homocysteine hydrolase, key enzyme of methylation metabolism, regulates phosphatidylcholine synthesis and triacylglycerol homeostasis in yeast: implications for homocysteine as a risk factor of atherosclerosis. *J Biol Chem*, 283(35):23989–99, Aug 2008.

Raimund Mannhold, Hugo Kubinyi, and Gerd Folkers. *Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective*, volume 22 of *Methods and Principles in Medicinal Chemistry*, chapter Preface. Wiley-VCH, 1 edition, 2004.

Gerard Manning. Introduction to Kinases, December 2006. URL `http://kinase.com/wiki/index.php/Introduction_to_Kinases`.

Gerard Manning. Protein Kinases: Introduction, N/A. URL `http://www.cellsignal.com/common/content/content.jsp?id=kinases`.

Robert Menard. Medicine: knockout malaria vaccine? *Nature*, 433(7022):113–4, Jan 2005.

Phelelani T Mpangase, Michal J Szolkiewicz, Misha le Grange, Jeanre H Smit, Pieter B Burger, and Fourie Joubert. Discovery-2: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. *Malar J*, 12:116, 2013.

Masayuki Nakanishi. S-adenosyl-L-homocysteine hydrolase as an attractive target for antimicrobial drugs. *Yakugaku Zasshi*, 127(6):977–82, Jun 2007.

Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat, and Peter Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–54, Nov 2004.

Leticia Orti, Rodrigo J Carbajo, Ursula Pieper, Narayanan Eswar, Stephen M Maurer, Arti K Rai, Ginger Taylor, Matthew H Todd, Antonio Pineda-Lucena, Andrej Sali, and Marc A Marti-Renom. A kernel for open source drug discovery in tropical diseases. *PLoS Negl Trop Dis*, 3(4):e418, 2009.

Arnab Pain and Christiane Hertz-Fowler. *Plasmodium* genomics: latest milestone. *Nat Rev Microbiol*, 7(3):180–1, Mar 2009.

Hui-Lin Pan, Zi-Zhen Wu, Hong-Yi Zhou, Shao-Rui Chen, Hong-Mei Zhang, and De-Pei Li. Modulation of pain transmission by G-protein-coupled receptors. *Pharmacol Ther*, 117(1):141–61, Jan 2008.

Deepak Perumal, Chu Sing Lim, and Meena K Sakharkar. A Comparative Study of Metabolic Network Topology between a Pathogenic and a Non-Pathogenic Bacterium for Potential Drug Target Identification. *Summit on Translat Bioinforma*, 2009:100–4, 2009.

Sharangdhar S Phatak and Shuxing Zhang. A novel multi-modal drug repurposing approach for identification of potent ACK1 inhibitors. *Pac Symp Biocomput*, pages 29–40, 2013.

Richard Pink, Alan Hudson, Marie-Annick Mouries, and Mary Bendig. Opportunities and challenges in antiparasitic drug discovery. *Nat Rev Drug Discov*, 4(9):727–40, Sep 2005.

Gurpur M. Prabhu. Computer Architecture Tutorial, 2009. URL `http://www.cs.iastate.edu/~prabhu/Tutorial/title.html`.

Syed Asad Rahman and Dietmar Schomburg. Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. *Bioinformatics*, 22(14):1767–74, Jul 2006.

Bilachi S Ravindranath, Venkatappa Krishnamurthy, Venkatarangaiah Krishna, and Kahale Bommaiah Lingaiah Vasudevanayaka. *In silico* analyses of metabolic pathway and protein interaction network for identification of next gen therapeutic targets in *Chlamydophila pneumoniae*. *Bioinformation*, 9(12):605–9, 2013.

Philip J Rosenthal. Antimalarial drug discovery: old and new approaches. *J Exp Biol*, 206(Pt 21): 3735–44, Nov 2003.

Mirjam Schunk, Wondimagegn P Kumma, Isabel Barreto Miranda, Maha E Osman, Susanne Roewer, Abraham Alano, Thomas Loscher, Ulrich Bienzle, and Frank P Mockenhaupt. High prevalence of drug-resistance mutations in *Plasmodium falciparum* and *Plasmodium vivax* in southern Ethiopia. *Malar J*, 5:54, 2006.

Govindan Subramanian, Adnan M M Mjalli, and Michael E Kutz. Integrated approaches to perform *in silico* drug discovery. *Curr Drug Discov Technol*, 3(3):189–97, Sep 2006.

J Tao, P Wendler, G Connelly, A Lim, J Zhang, M King, T Li, J A Silverman, P R Schimmel, and F P Tally. Drug target validation: lethal infection blocked by inducible peptide. *Proc Natl Acad Sci U S A*, 97(2):783–6, Jan 2000.

Leslie W Tari, Xiaoming Li, Michael Trzoss, Daniel C Bensen, Zhiyong Chen, Thanh Lam, Junhu Zhang, Suk Joong Lee, Grayson Hough, Doug Phillipson, Suzanne Akers-Rodriguez, Mark L Cunningham, Bryan P Kwan, Kirk J Nelson, Amanda Castellano, Jeff B Locke, Vickie Brown-Driver, Timothy M Murphy, Voon S Ong, Chris M Pillar, Dean L Shinabarger, Jay Nix, Felice C Lightstone, Sergio E Wong, Toan B Nguyen, Karen J Shaw, and John Finn. Tricyclic GyrB/ParE (TriBE) inhibitors: a new class of broad-spectrum dual-targeting antibacterial agents. *PLoS One*, 8(12):e84409, 2013.

TDR. User's Manual for TDRtargets.org, January 2011. URL `http://tdrtargets.org/manual`.

TechTarget. Wrapper, April 2005. URL `http://searchsoa.techtarget.com/definition/wrapper`.

TechTarget. Web services (application services), March 2007. URL `htthttp://searchsoa.techtarget.com/definition/Web-services`.

David Toomey, Heinrich C Hoppe, Marian P Brennan, Kevin B Nolan, and Anthony J Chubb. Genomes2Drugs: identifies target proteins and lead drugs from proteome data. *PLoS One*, 4(7):e6195, 2009.

Leah S Torrie, Susan Wyllie, Daniel Spinks, Sandra L Oza, Stephen Thompson, Justin R Harrison, Ian H Gilbert, Paul G Wyatt, Alan H Fairlamb, and Julie A Frearson. Chemical validation of trypanothione synthetase: a potential drug target for human trypanosomiasis. *J Biol Chem*, 284(52):36137–45, Dec 2009.

Mark A Travassos and Miriam K Laufer. Resistance to antimalarial drugs: molecular, pharmacologic, and clinical considerations. *Pediatr Res*, 65(5 Pt 2):64R–70R, May 2009.

T Triglia, J G Menting, C Wilson, and A F Cowman. Mutations in dihydropteroate synthase are responsible for sulfone and sulfonamide resistance in plasmodium falciparum. *Proc Natl Acad Sci U S A*, 94(25):13944–9, Dec 1997.

UniProtFAQ. Frequently Asked Questions, N/A. URL `http://www.uniprot.org/faq`.

UniProtManual. UniProt Knowledgebase Swiss-Prot Protein Knowledgebase TrEMBL Protein Database User Manual, July 2014. URL `http://web.expasy.org/docs/userman.html`.

Santiago Vilar and Stefano Costanzi. Predicting the biological activity through QSAR analysis and docking-based scoring. *Methods Mol Biol*, 2012.

David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res*, 34(Database issue):D668–72, Jan 2006.

Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–54, Jun 2010.

Wu-Lung R Yang, Yu-En Lee, Ming-Huang Chen, Kun-Mao Chao, and Chi-Ying F Huang. *In silico* drug screening and potential target identification for hepatocellular carcinoma using Support Vector Machines based on drug screening result. *Gene*, 518(1):201–8, Apr 2013.

Iwei Yeh, Theodor Hanekamp, Sophia Tsoka, Peter D Karp, and Russ B Altman. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome Res*, 14(5):917–24, May 2004.

Yongyuth Yuthavong, Bongkoch Tarnchompoo, Tirayut Vilaivan, Penchit Chitnumsub, Sumalee Kamchonwongpaisan, Susan A Charman, Danielle N McLennan, Karen L White, Livia Vivas, Emily Bongard, Chawanee Thongphanchang, Supannee Taweechai, Jarunee Vanichtanankul, Roonglawan Rattanajak, Uthai Arwon, Pascal Fantauzzi, Jirundon Yuvaniyama, William N Charman, and David Matthews. Malarial dihydrofolate reductase as a paradigm for drug development against a resistance-compromised target. *Proc Natl Acad Sci U S A*, 109(42):16823–8, Oct 2012. doi: 10.1073/pnas.1204556109.