

Integrating protein annotations for the *in silico*
prioritization of putative drug target proteins in
malaria

by

Phelelani Mpangase

Submitted in partial fulfillment of 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

November 2012

Declaration

I, Phelelani Thokozani Mpangase, declare that the thesis/dissertation, which I hereby submit for the degree *Magister Scientiae* at the University of Pretoria, is my own work and not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: Date:

Acknowledgements

I would like to thank the following people for their contribution towards the completion of this thesis:

- Professor Fourie Joubert for his professional support and guidance with the project and writing of this thesis.
- My parents and brother for their love and support in all my studies and the decisions I make.
- Jeanré Smith and Michal Szolkiewicz who I worked closely with in my project.
- Oliver Bezuidt and my colleagues at the Bioinformatics and Computational Biology Unit of the University of Pretoria for all their help and advice.
- John Overington and Louisa Bellis for making the internship at the European Bioinformatics Institute (EBI) possible.
- Kazuyoshi Ikeda and the ChEMBL team for their help with the druggability data.
- The Department of Science & Technology (DST) of South Africa, National Research Foundation (NRF) and the University of Pretoria for the funding which made it possible to complete my studies.

Summary

Current anti-malarial methods have been effective in reducing the number of malarial cases. However, these methods do not completely block the transmission of the parasite. Research has shown that repeated use of the current anti-malarial drugs, which include artemisinin-based drug combinations, might be toxic to humans. There have also been reports of an emergence of artemisinin-resistant parasites. Finding anti-malarial drugs through the drug discovery process takes a long time and failure results in a great financial loss. The failure of drug discovery projects can be partly attributed to the improper selection of drug targets. There is thus a need for an effective way of identifying and validating new potential malaria drug targets for entry into the drug discovery process.

The availability of the genome sequences for the *Plasmodium* parasite, human host and the *Anopheles* mosquito vector has facilitated post-genomic studies on malaria. Proper utilization of this data, in combination with computational biology and bioinformatics techniques, could aid in the *in silico* prioritization of drug targets. This study was aimed at extensively annotating the protein sequences from the *Plasmodium* parasites, *H. sapiens* and *A. gambiae* with data from different online databases in order to create a resource for the prioritization of drug targets in malaria. Essentiality, assay feasibility, resistance, toxicity, structural information and druggability were the main target selection criteria which were used to collect data for protein annotations. The data was used to populate the Discovery resource (<http://malport.bi.up.ac.za/>) for the *in silico* prioritization of potential drug targets.

A new version of the Discovery system, Discovery 2.0 (<http://discovery.bi.up.ac.za/>), has been developed using Java. The system contains new and automatically updated data as well as improved functionalities. The new data in Discovery 2.0 includes UniProt accessions, gene ontology annotations from the UniProt-GOA project, pathways from Reactome and Malaria Parasite Metabolic Pathways databases, protein-protein interactions data from

IntAct as well as druggability data from the DrugEBlity resource hosted by ChEMBL. Users can access the data by searching with a protein identifier, UniProt accession, protein name or through the advanced search which lets users filter protein sequences based on different protein properties. The results are organized in a tabbed environment, with each tab displaying different protein annotation data.

A sample investigation using a previously proposed malarial target, S-adenosyl-L-homocysteine hydrolase, was carried out to demonstrate the different categories of data available in Discovery 2.0 as well as to test if the available data is sufficient for assessment and prioritization of drug targets. The study showed that using the annotation data in Discovery 2.0, a protein can be assessed, in a species comparative manner, on the potential of being a drug target based on the selection criteria mentioned here. However, supporting data from literature is also needed to further validate the findings.

Contents

Declaration	i
Aknowledgements	ii
Summary	iii
Table of Contents	viii
List of Figures	ix
List of Tables	xi
List of Algorithms	xii
Abbreviations	xiii
Chapter 1: Introduction	1
1.1 Target discovery	3
1.1.1 System-based target discovery	4
1.1.2 Molecular-based targets discovery	5
1.2 Genomic sequencing	6
1.3 Target assessment	8
1.3.1 Essentiality	8
1.3.2 Assay feasibility	10
1.3.3 Resistance	12
1.3.4 Toxicity	13
1.3.5 Structural information	15
1.3.6 Druggability	17
1.4 Problem statement	19
1.5 Aims	20
Chapter 2: Methods	21
2.1 Introduction	21

2.2	Protein sequences and function	22
	<i>PlasmoDB</i>	22
	<i>VectorBase</i>	23
	<i>Ensembl</i>	23
	<i>UniProt</i>	24
	<i>UniProt-GOA</i>	25
	<i>InterPro</i>	25
	2.2.1 Obtaining protein sequences	26
	2.2.2 Functional annotation	27
2.3	Orthology	27
	<i>OrthoMCL</i>	27
	2.3.1 Assignment of sequences to orthologous groups using OrthMCL	29
	2.3.2 Multiple sequence alignment using T-coffee	29
2.4	Structural information	29
	<i>PDB</i>	29
	<i>Modbase</i>	30
	2.4.1 BLAST search against PDB database	31
	2.4.2 Predicted MODBASE structures	31
2.5	Metabolic pathways and enzyme information	32
	<i>KEGG</i>	32
	<i>MPMP</i>	32
	<i>Reactome</i>	34
	<i>ExpASy</i>	35
	<i>BRENDA</i>	36
	2.5.1 Metabolic pathway assignment	36
	2.5.2 EC number assignment linking to databases	37
2.6	Protein-protein interactions	37
	2.6.1 Assignment of protein-protein interactions	38
2.7	Druggability	38
	<i>DrugEBility</i>	38

2.7.1	BLAST search against DrugEBIity database	39
2.8	Discussion	39
Chapter 3: Results and discussion		41
3.1	Introduction	41
3.2	The Discovery 2.0 web-interface	42
3.2.1	Summary	45
3.2.2	Function	46
3.2.3	Gene Ontology	47
3.2.4	Orthology	47
3.2.5	Metabolic pathways	48
3.2.6	Structure	49
3.2.7	Interactions	50
3.2.8	Druggability	51
3.3	The annotation data in Discovery 2.0	52
3.4	Case studies on Discovery 2.0	57
3.4.1	Protein kinase	59
3.4.2	G protein-coupled receptor	62
3.4.3	Peptidase	66
3.4.4	Aminopeptidase	70
3.4.5	Dehydrogenase	72
3.5	Assessment of a protein target using Discovery 2.0	75
	<i>Summary</i>	76
	<i>Function</i>	76
	<i>Gene ontology</i>	77
	<i>Orthology</i>	77
	<i>Structures</i>	77
	<i>Metabolic pathways</i>	78
	<i>Interactions</i>	78
	<i>Druggability</i>	78
3.5.1	Essentiality	79

3.5.2	Assay feasibility	81
3.5.3	Resistance	82
3.5.4	Toxicity	83
3.5.5	Structural information	84
3.5.6	Druggability	87
3.6	Prioritization of potential drug targets in malaria using Discovery 2.0	89
3.7	Discussion	91
	Chapter 4: Concluding discussion	93
	Bibliography	98

List of Figures

1.1	Summary of the methods used in the two different approaches to target discovery	3
1.2	Choke-point analysis	10
1.3	Reaction catalyzed by DHOD	11
1.4	The 424 amino acid <i>Pf</i> CRT transmembrane protein encoded by the 13-exon <i>pfCRT</i> gene	13
1.5	Regulation and expression of human and <i>Plasmodium</i> DHFR	14
1.6	Docking of WR99210 analogues to mutant DHFR	16
2.1	Clustering of orthologs using the OrthoMCL algorithm	28
2.2	Nitrogen metabolism pathway for the <i>Plasmodium</i> parasite	34
3.1	Discovery 2.0 home page	42
3.2	Advanced search	43
3.3	Reaction catalyzed by dUTPase.	43
3.4	Summary tab	45
3.5	Predicted functions tab	46
3.6	Gene Ontology tab	47
3.7	Orthology tab	48
3.8	Metabolic pathways tab	49
3.9	Crystal structures tab	50
3.10	Interactions tab	51
3.11	Druggability tab	51
3.12	Genome annotations	53
3.13	MODBASE statistics for the modelled genomes.	56
3.14	Search for proteins by EC numbers in PlasmoDB	57

3.15	An advanced search to identify a protein sequence belonging to the protein kinase superfamily in <i>P. falciparum</i>	60
3.16	An advanced search in Discovery 2.0 for identifying a GPCR protein sequence in <i>P. falciparum</i>	64
3.17	An advanced search carried out in Discovery 2.0 to identify aspartic proteases sequences in <i>P. falciparum</i>	67
3.18	An advanced search in Discovery 2.0 for identifying a peptidase in <i>P. falciparum</i> .	70
3.19	An advanced search in Discovery 2.0 carried out to identify the <i>P. falciparum</i> enzyme DHOD (<i>PfDHOD</i>).	73
3.20	Hydrolysis of S-adenosyl-L-homocysteine to adenosine and L-homocysteine. . . .	76
3.21	Analysis of <i>PfSAHH</i> in PlasmoDB genome browser	79
3.22	Methionine and polyamine metabolism	80
3.23	Advanced search to identify proteins with the same or similar function to <i>PfSAHH</i> .	82
3.24	Crystal structure of the tetrameric <i>PfSAHH</i> enzyme	85
3.25	Crystal structure of the <i>PfSAHH</i> subunit	86
3.26	Active site of <i>PfSAHH</i>	86
3.27	Undruggable sites identified at the known binding sites on the two <i>PfSAHH</i> domains	87
3.28	Summary of the druggability calculations for <i>PfSAHH</i>	88

List of Tables

1.1	Online resources relevant to malaria	7
2.1	KEGG databases	33
3.1	Different types of filters available on the advanced search of Discovery 2.0.	58
3.2	<i>P. falciparum</i> MO15-related protein kinase (PF10_0141) annotation summary.	61
3.3	<i>P. falciparum</i> G-protein coupled receptor (PFE1265w) annotation summary.	65
3.4	<i>P. falciparum</i> plasmepsin I (PF14_0076) annotation summary.	69
3.5	<i>P. falciparum</i> M17 leucyl aminopeptidase (PF14_0439) annotation summary.	71
3.6	<i>PfDHOD</i> (PFF0160c) annotation summary.	74
3.7	Summary of the InterPro signatures matching the <i>PfSAHH</i> protein sequence.	77
3.8	Summary for the assessment of <i>PfSAHH</i> as a drug target.	90

List of Algorithms

1.1	Model for calculating druggability	18
-----	--	----

Abbreviations

API	Application programming interface
aPK	Atypical protein kinases
BNL	Brookhaven National Laboratories
BRENDA	Braunschweig Enzyme database
CAP	Community annotation pipeline
CDK	Cyclin-dependent protein kinase
CoQ	Co-enzyme Q/Ubiquinone
CoQH ₂	Ubiquinol
DCIP	2,6-dichlorophenol-indophenol
DDT	Dichloro-diphenyl-trichloroethane
DHF	Dihydrofolate
DHFR-TS	Dihydrofolate reductase-thymidylate synthase
DHOD	Dihydroorotate dehydrogenase
DOPE	Discrete Optimized Protein Energy
dTMP	Deoxythymidine monophosphate
dTTP	Deoxythymidine triphosphate
dUMP	Deoxyuridine monophosphate

dUTP	Deoxyuridine triphosphate
dUTPase	deoxyuridine 5'-triphosphate nucleotidohydrolase
EBI	European Bioinformatics Institute
EC	Enzyme commission
ePK	Eukaryotic protein kinases
ExPASy	Expert Protein Analysis System
FMN	Flavin mononucleotide
FTP	File transfer protocol
GO	Gene ontology
GPCR	G protein-coupled receptors
HPLC	High-pressure liquid chromatography
HsdUTPase	<i>H. sapiens</i> dUTPase
HsSAHH	<i>H. sapiens</i> S-adenosyl-L-homocysteine hydrolase
HTS	High-throughput screening
IMEx	International Molecular Interaction Exchange consortium
IRS	Indoor residual spraying
IUBMB	International Union of Biochemistry and Molecular Biochemistry
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAP-3	Leucine aminopeptidase 3
MCL	Markov Clustering algorithm
MPMP	Malaria Parasite Metabolic Pathways

MR	Molecular replacement
NAD	Nicotinamide adenine dinucleotide
NAM	Noraristeromycin
NMR	Nuclear magnetic resonance
PDB	Protein Data Bank
PfCRT	Chloroquine Resistance Transporter
PfDHOD	<i>P. falciparum</i> Dihydroorotate dehydrogenase
PfdUTPase	<i>P. falciparum</i> dUTPase
PfSAHH	<i>P. falciparum</i> S-adenosyl-L-homocysteine hydrolase
PSI-MI	Proteomics Standard Initiative-Molecular Interaction
REX3	Ring-exported protein
RNAi	RNA interference
RSCB	Research Collaboratory for Structural Bioinformatics
SAH	S-adenosyl-L-homocysteine
SAM	S-adenosyl-methylthionine
SIB	Swiss Institute of Bioinformatics
THF	Tetrahydrofolate
UniMES	UniProt Metagenomic and Environmental Sequences database
UniParc	UniProt archive
UniProt	Universal Protein Resource
UniProtKB	The UniProt Knowledgebase
UniRef	UniProt Reference Clusters
vHTS	Virtual High-throughput Screening

Chapter 1

Introduction

Malaria continues to be amongst the major causes of death in developing tropical and sub-tropical countries, claiming lives of about 2.7 - 3 million people annually (Aurrecoechea *et al.*, 2009; Fatumo *et al.*, 2009). The disease is caused by the parasites of the *Plasmodium* species, of which *Plasmodium falciparum* is the major cause of malaria deaths in humans; *Plasmodium vivax*, *Plasmodium malariae* and *Plasmodium ovale* are amongst other *Plasmodium* parasites that can infect humans (Gardner *et al.*, 2002; Aurrecoechea *et al.*, 2009). The malaria causing parasite is transmitted to humans by an *Anopheles gambiae* female mosquito during a blood meal (Holt *et al.*, 2002). With the mosquito as a vector and humans as hosts for the malaria-causing parasite, the parasite lives a complicated life and depends on both the vector and host mechanisms for survival.

Even though the life of the *Plasmodium* parasite is complicated, the disease it causes in humans is curable and can be prevented (Joubert *et al.*, 2009). Early methods for malaria eradication were mainly based on vector control. Larvicidal methods were used in the late 19th century, where the *Anopheles* mosquitoes were disrupted at their breeding stages to try and reduce the number of mosquito vectors transmitting the parasite to humans (Bruce-Chwatt, 1981). Dichloro-diphenyl-trichloroethane (DDT) was used as an insecticide to control the *Anopheles* mosquitoes in the 1940's, but failed due to its toxicity to the environment and the emergence of DDT resistant *Anopheles* mosquitoes (Bruce-Chwatt, 1981). However, in parts of Southern Africa, indoor residual spraying (IRS) with DDT is still used as a method for vector control due to its low-cost and effectiveness (Sadasivaiah *et al.*, 2007; Bornman *et al.*, 2010; Bouwman *et al.*, 2011).

Apart from vector controlling methods, patients infected with the malaria parasite were treated with chloroquine and quinine drugs. However, these drugs also failed due to the resistance of *Plasmodium* parasites to drugs (Cowman *et al.*, 1994; Djimdé *et al.*, 2001; Sidhu *et al.*, 2002). Recent reports show that the deaths caused by malaria are estimated to have decreased. This decrease in malaria deaths is mainly due to the current anti-malarial methods being employed. These methods include combination therapy, where current drugs like artemisinin (and its derivatives) are combined with other new drugs to prolong the drug use and reduce parasite resistance (Nosten *et al.*, 2000; Adjuik *et al.*, 2004; Longo *et al.*, 2006; Dondorp *et al.*, 2009). Using insecticide-treated bed nets and IRS with DDT are other methods used to decrease malarial infection (Sadasivaiah *et al.*, 2007; Oxborough *et al.*, 2008; Bornman *et al.*, 2010; Bouwman *et al.*, 2011).

However, even though artemisinin-based combination therapy has been successful in the fight against malaria, studies have shown that the repeated use of artemisinin can be toxic to humans as it is in mouse models (Afonso *et al.*, 2006). Noedl *et al.* (2008) and Dondorp *et al.* (2009) also reported an emergence of artemisinin-resistance parasites. There is thus a need to find new potential protein targets for drug design which will replace current anti-malarial methods. Drug discovery, however, is a lengthy and very expensive process. It involves the identification of a protein target (which must be essential for the progression of the disease under study), validating that the protein target is important to the progression of the disease, identifying leads active against the target and optimizing them for human safety before they can finally be used by the public (Chen and Chen, 2008). Failure to identify essential protein targets for drug discovery can result in a great financial loss.

The availability of the complete genomic sequences of human (Venter *et al.*, 2001), *P. falciparum* (Gardner *et al.*, 2002) and *A. gambiae* (Holt *et al.*, 2002; Sharakhova *et al.*, 2007) has facilitated post-genomic studies on malaria. Bioinformatics and Computational Biology techniques utilize the genomic sequences to predict the biological roles and properties of proteins encoded in the genome (Chen *et al.*, 2006; Kanehisa *et al.*, 2006, 2008). With the availability of such information, the risk of selecting targets that may fail in the drug discovery process may be greatly reduced if the information is utilized properly to predict and prioritize protein targets. In this chapter, ways of selecting and prioritizing drug targets *in silico* for entry into

the drug discovery process using post-genomic data and resources available online are discussed, with the emphasis being on resources relevant to malaria.

1.1 Target discovery

A target can be any biological entity or phenomenon that is crucial to a particular disease and its progression. Chen and Chen (2008) define a target as being anything from a gene, protein domain, protein, organelle or a biological process. The drug discovery process aims at identifying such biological entities and finding drugs that will be able to modulate them, thus inhibiting the progression of a disease.

Target identification is the first and most crucial step in the drug discovery process. Failure to identify targets that are crucial to the progression of a disease can result in a great financial and time loss. It is thus important to make informed decisions using available information on a particular target before it is entered into the drug discovery pipeline. Two major approaches can be distinguished in the area of target discovery, the “system-based” (Nolan *et al.*, 2000) and “molecular-based” approaches (Zhang and Rathod, 2002; Wu and Ding, 2007). Figure 1.1 summarizes the two different approaches to target discovery. In the following sections (Section 1.1.1 and 1.1.2), these two approaches to target discovery are discussed, as well as their

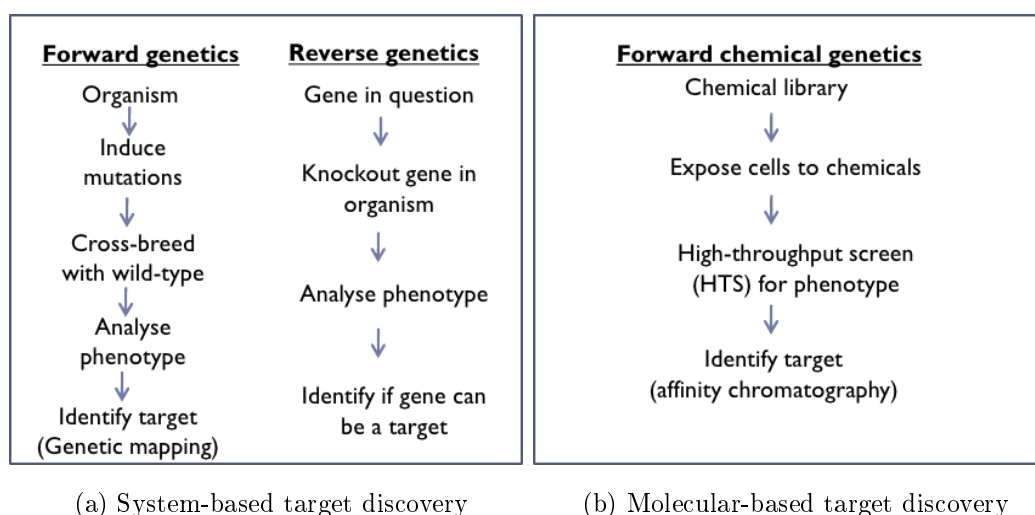


Figure 1.1: **Summary of the methods used in the two different approaches to target discovery.** The system-based approach (a) to target discovery incorporates forward and reverse genetics whilst the molecular-based approach (b) is composed mainly of forward chemical genetics (Lindsay, 2003; Wu and Ding, 2007).

advantages and disadvantages.

1.1.1 System-based target discovery

The systems-based approach to target discovery (Figure 1.1a) focuses on the identification of targets through the studying of the disease in a whole organism (Nolan *et al.*, 2000). This is a more traditional approach to target discovery. The approach uses different *in vivo* phenotype-driven techniques in model organisms for the identification of targets. Two techniques are used in system-based target discovery, “forward genetics” and “reverse genetics” (Stockwell *et al.*, 1999).

Forward genetics is aimed at target identification through alteration of a phenotype by inducing mutations in a disease model organism, which is then cross-bred with a wild-type organism and phenotypic screening is performed on off-springs to identify relevant mutants (gain or loss of function) (Stockwell *et al.*, 1999; Nolan *et al.*, 2000; de Angelis *et al.*, 2000). Once the mutants are identified, the gene responsible for the resulting phenotype is then identified by genetic mapping. The advantage of this technique is that no prior knowledge about the gene is needed (de Angelis *et al.*, 2000; Nolan *et al.*, 2000). However, the major issues of this technique is the time it takes to produce mutants organism and to do the analysis as well as the genetic differences between the model organism and humans (Stockwell *et al.*, 1999).

Reverse genetics (Figure 1.1a) on the other hand involves identification of a target through studying of phenotypic effects of gene manipulation in an disease model organism (Stockwell *et al.*, 1999). The gene being manipulated may be a molecule of unknown function which is obtained by sequencing. Gene knockouts, insertional mutagenesis and RNA interference (RNAi) are some of the techniques used for gene manipulation used to identify the role played by a gene in the progression of a disease in the model organism (Blandin *et al.*, 2002). Reverse genetics is a much more rapid and specific method of identifying essential genes in disease model organism, however, the inability to produce viable organism or phenotype is a major limitation to this approach.

1.1.2 Molecular-based targets discovery

The molecular-based approach (Figure 1.1b) aims at studying cells involved in the disease rather than the whole organism to identify targets (Patel *et al.*, 2008; Buchholz *et al.*, 2011). A number of techniques are used in molecular-based target discovery, which include genomics and proteomics. Genomics is aimed at target identification through comparison of gene expression levels in normal and disease tissues, whereas proteomics compares protein expression. As with system-based target discovery, molecular-based target discovery also encompasses forward genetics as one of the techniques for target identification, but on a cellular level. Instead of using gene manipulation, forward genetics in molecular-based target discovery makes use of small molecules to inhibit or modulate a protein, which affects certain biological processes and thereby producing certain phenotype in the cells (Wu and Ding, 2007). This technique is called “forward chemical genetics”.

In forward chemical genetics, a library of small molecules are screened for their ability to inhibit cellular processes by interacting with certain proteins in the cell. A high-throughput screening (HTS) assay for the particular cells in question is needed to identify the effects caused by the small molecules on the cells. If small molecules are able to induce changes in the cells (based on the phenotype observed in HTS), the protein target which it is active against needs to be identified using biochemical affinity-based methods (Wu and Ding, 2007).

Three components are required for forward chemical genetics: a) a collection (library) of small molecules that will be screened for their ability to inhibit cellular processes, b) a HTS for the effects caused by the small molecules on cells and c) a method to identify the protein targets that small molecules are active against (Weisman *et al.*, 2006; Wu and Ding, 2007; Patel *et al.*, 2008). This technique has an advantage over other methods used for target discovery as it not only identifies potential targets, but it also gives potential leads for drug development by providing minimal drug-like properties for the active molecule like cellular activity, solubility and cell permeability (Baniecki *et al.*, 2007; Buchholz *et al.*, 2011). However, identification of targets that small molecules are active against still remains an issue, especially if the targets are available in low concentrations in the cell (Evans *et al.*, 2005).

1.2 Genomic sequencing

In 2001, the complete human genome sequence generated by whole-genome shotgun sequencing method was released (Venter *et al.*, 2001). This was followed by the subsequent release of the genome sequences of the *A. gambiae* (Holt *et al.*, 2002) and *P. falciparum* (Gardner *et al.*, 2002) in 2002. Even though the sequencing of these three organisms had its challenges, the availability of the genome sequences facilitated post-genomic studies especially in understanding the underlying mechanisms of many parasites that cause diseases in human. For example, using the available *P. falciparum* genomic sequencing data, Yeh *et al.* (2004) were able to computationally analyze *P. falciparum* metabolic networks in an effort to identify potential drug targets. Using their method, they were able to identify three clinically proven malaria drug targets (dihydrofolate reductase, dihydropteroate synthase and 1-deoxy-D-xylulose) and proposed 24 possible drug targets for malaria (Yeh *et al.*, 2004).

The *P. falciparum* parasite's nuclear genome is packaged into 14 chromosomes (22.8 megabases), with an approximation of 5 300 protein-coding genes (Gardner *et al.*, 2002). The high (A+T) content and the larger gene length in the *P. falciparum* genome reported by Gardner *et al.* (2002) was the cause of limited success in sequence similarity searching when trying to assign function to genes. About 60% of the predicted 5 268 proteins in *P. falciparum* did not have any similarity to proteins in other organisms. The results of the ongoing *P. falciparum* genome annotation, as well as other *Plasmodium* species, are available on PlasmoDB (<http://plasmodb.org>), a functional genomic database for *Plasmodium* species (Gardner *et al.*, 2002; Kissinger *et al.*, 2002; Aurrecochea *et al.*, 2009)). Apart from PlasmoDB, other resources utilizing Bioinformatics and Computational Biology to increase the knowledge on the *Plasmodium* parasite are available (Table 1.1)

Databases dedicated to the annotation of the human genome (<http://ensembl.org>) (Hubbard *et al.*, 2009) as well as *A. gambiae* vector (<http://www.vectorbase.org>) (Lawson *et al.*, 2009) also exist. Proper integration of all the resources available for the three species involved in malaria can be beneficial in understanding the mechanisms of the disease and identification of possible targets *in silico* that may be used for drug discovery. The TDR targets databases (<http://tdrtargets.org>) is one of the resource available for the prioritization of drug targets for neglected disease (Agüero *et al.*, 2008). It encompasses protein druggability predictions

Table 1.1: **Online resources relevant to malaria**

Database	Description
PlasmoDB	Main repository for all sequenced <i>Plasmodium</i> species. Contains annotated genomes, transcript expression data, protein expression data, GO annotation and protein localization. (http://plasmodb.org/) (Aurrecochea <i>et al.</i> , 2009).
PlasmoCyc	A pathway/genome database (PGDB) for <i>P. falciparum</i> . The database contains a metabolic network predicted computationally using the <i>P. falciparum</i> genome. (http://plasmocyc.stanford.edu/) (Yeh <i>et al.</i> , 2004).
MPMP	Malaria Parasite Metabolic Pathways, a database for manually reconstructed and curated metabolic pathways for the intra-erythrocytic phase of <i>P. falciparum</i> . (http://sites.huji.ac.il/malaria/) (Ginsburg, 2006).
PlasmoDraft	Database containing Gene Ontology (GO) annotations for <i>P. falciparum</i> predicted from post-genomic data. (http://atgc.lirmm.fr/PlasmoDraft/) (Br�h�lin <i>et al.</i> , 2008).
MalVac	Database containing potential malarial vaccine candidates. (http://malvac.igib.res.in/) (Chaudhuri <i>et al.</i> , 2008).
Discovery	Database that incorporates protein information for <i>Plasmodium</i> , <i>H. sapiens</i> and <i>A. gambiae</i> as well as chemical information for malarial lead and target selection. (http://malport.bi.up.ac.za/) (Joubert <i>et al.</i> , 2009).
TDR Targets	Database for identification and prioritization of drug targets for neglected tropical diseases. (http://tdrtargets.org/) (Ag�ero <i>et al.</i> , 2008).
TDI kernel	An open source drug discovery kernel for large-scale prioritization of targets, identification of binding sites for small molecules and lead identification. (http://tropicaldisease.org/) (Ort� <i>et al.</i> , 2009).

(druggability discussed in Section 1.3.6) together with genetic, biochemical and pharmacological data for disease causing pathogens. This resource also allows weighting of this data according to the users preference in order to prioritize potential drug target candidates.

Another resource, Discovery (<http://malport.bi.up.ac.za>), hosted at the University of Pretoria is also aimed at bringing together all available post-genomic data relevant to malaria (Joubert *et al.*, 2009). This resource is unique in that it not only brings together data from the malaria parasites but also integrates chemical data as well as data available for the host (*Homo sapiens*) and the vector (*A. gambiae*). Inclusion of chemical data in Discovery also helps in selecting chemical leads. Such integration or mining of post-genomic data as seen in TDR targets database and Discovery allows comparison and rational selection of candidate targets for drug discovery. It is thus important to define the most important *in silico* target assessment

criterion so that valuable information can be integrated in such databases.

1.3 Target assessment

The drug discovery process can be facilitated by applying *in silico* target discovery. As mentioned in Section 1.1, target identification and validation using traditional and current methods can take a very long time and can sometimes fail due to the inability to produce viable organisms or phenotype and also by the inability to isolate the target that a small molecule is active against. With *in silico* target discovery, the target assessment criterion can be defined and this can be used to gather and mine the data that will be useful in assessing a target. However, *in vivo* and *in vitro* methods are still required to validate the targets. In developing the TDR targets database, Agüero *et al.* (2008) defined the most important target assessment criteria, which may be summarized into six main categories i.e., a) essentiality, b) assay feasibility, c) resistance, d) toxicity, e) structural information and f) druggability. In the following sections (Section 1.3.1 - 1.3.6), these target assessment criteria are discussed.

1.3.1 Essentiality

For a protein to be considered as a target, it must be shown to be involved in a biochemical process or pathway that is crucial for the progression of a disease (Fatumo *et al.*, 2009; Doyle *et al.*, 2010). If such proteins are inhibited, the parasite that causes the disease is unable to survive and thus the symptoms of the disease are reversed. Data about the essentiality of gene products can be obtained in various ways. Analyzing literature (automated or manual curation) for experimentally validated essential genes (e.g. gene knockouts, RNAi) in an organism is one way of identifying essentiality (Agüero *et al.*, 2008). However, not much essentiality data is published for some organisms due to the difficulty in performing gene knockouts in many parasites, including *Plasmodium* parasites (Doyle *et al.*, 2010).

Another method for identifying essentiality is through orthology. In this method, a gene product is said to be essential if it has a homologous gene that has been validated to be essential in model organisms (Agüero *et al.*, 2008; Doyle *et al.*, 2010). This method comes from the observation that essential genes are more likely to be conserved between species as compared to non-essential genes (Mushegian and Koonin, 1996; Curran and Ruvkun, 2007;

Doyle *et al.*, 2010). Assigning essentiality to proteins using this method can be risky as it is not always true that a validated essential orthologous gene in one organism is also essential in another organism (Agüero *et al.*, 2008).

Manually analyzing metabolic pathways in which a protein is involved in is another way of identifying essentiality. A number of pathway databases are available. These include the Kyoto Encyclopedia of Genes and Genomes (KEGG, <http://www.genome.jp/kegg/>), which is aimed at construction of pathways for different organisms using genomic and chemical information (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2012). PlasmoCyc and Malaria Parasite Metabolic Pathways (MPMP), accessible at <http://plasmocyc.stanford.edu/> and <http://sites.huji.ac.il/malaria/> respectively, are pathway databases that have been specifically constructed for the *Plasmodium* parasite. PlasmoCyc was constructed automatically using the annotated proteins of the *Plasmodium* parasite and a reference database of previously described pathways (Yeh *et al.*, 2004). MPMP on the other hand was manually reconstructed and curated using data from other pathway databases and websites (Ginsburg, 2006). Manually analyzing metabolic pathways for essential targets can be a very difficult task as one cannot come to a conclusions just by inspecting a pathway, additional information from literature is needed.

A more advanced method of identifying essentiality of gene products is through the *in silico* analysis of metabolic pathways, a method that has been used by Yeh *et al.* (2004) and Fatumo *et al.* (2009) for identification of potential drug targets in *P. falciparum*. In this method, a metabolic network of all metabolic reactions is created using metabolites i.e., two reactions are connected (neighbors) if the product of one reaction is the substrate of the other reaction (Fatumo *et al.*, 2009). Once a metabolic network has been created, “choke-point” analysis is done on the network using a defined algorithm. A “choke-point reaction” is a reaction that uniquely produces or consumes a metabolite in a metabolic network (Yeh *et al.*, 2004). Briefly, a reaction is deleted in the network and the ability of the network to produce downstream product in the absence of that reaction is analyzed (Figure 1.2).

If the downstream metabolites are produced, it means that other mechanisms or enzymes in the pathways are producing the metabolite, and thus the reaction is not essential. However, if there are no downstream metabolites produced after knocking out the choke-point reaction,

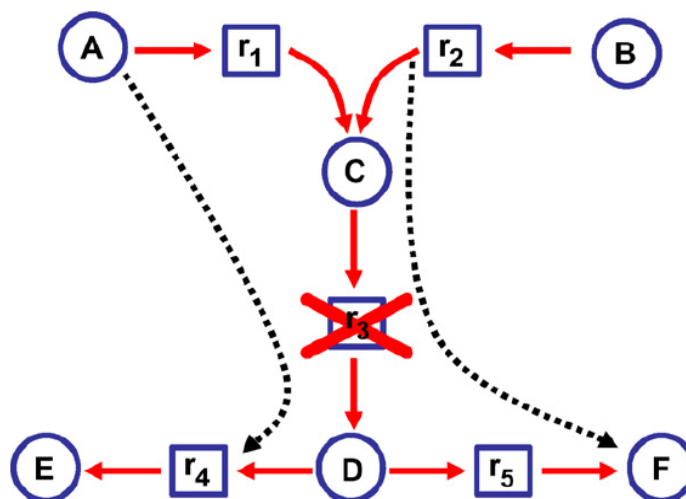


Figure 1.2: **Choke-point analysis.** If the metabolites E and F are produced in the absence of the knocked out (choke-point) reaction, then the reaction is not essential. [Adapted from Fatumo *et al.* (2009)].

then the reaction is essential and it may be concluded that the enzyme catalyzing that reaction is an essential target. The advantage to this method is that it does not only identify essential reactions and targets, it can also be used to identify potential resistance mechanisms.

1.3.2 Assay feasibility

Assay feasibility refers to the ability to readily perform an assay for a given target using available protocols and reagents (Agüero *et al.*, 2008). The ability to perform an assay on an identified target is important as it facilitates drug discovery when it comes to screening for chemical leads. Assays that will allow binding of small molecules to the target in molecular-based and cell-based HTS as well as detection of activity are needed. Developing such assays, however, is usually limited by the availability of a soluble recombinant target protein in large quantities; a protocol for cloning, expressing and purification of a target protein is required. A good example of how important availability of assay information is in drug discovery is seen in the work done by Patel *et al.* (2008).

Their work describes how they identified small molecules that are active against dihydroorotate dehydrogenase (DHOD), an attractive malaria target involved in the *de novo* pyrimidine biosynthetic pathway which catalyzes the oxidation of L-Dihydroorotate to orotate (Figure 1.3) (Patel *et al.*, 2008). Prior to HTS for identification of small molecules active against *P. falciparum* DHOD (*Pf*DHOD), they had to obtain the target protein in large quantities, pure and

soluble. Patel *et al.* (2008) achieved this by cloning a synthetic *Pf*DHOD gene into plasmid vector, then transforming *Escherichia coli* cells with the expression constructs. The target protein was then expressed and purified from the cell cultures and used for HTS.

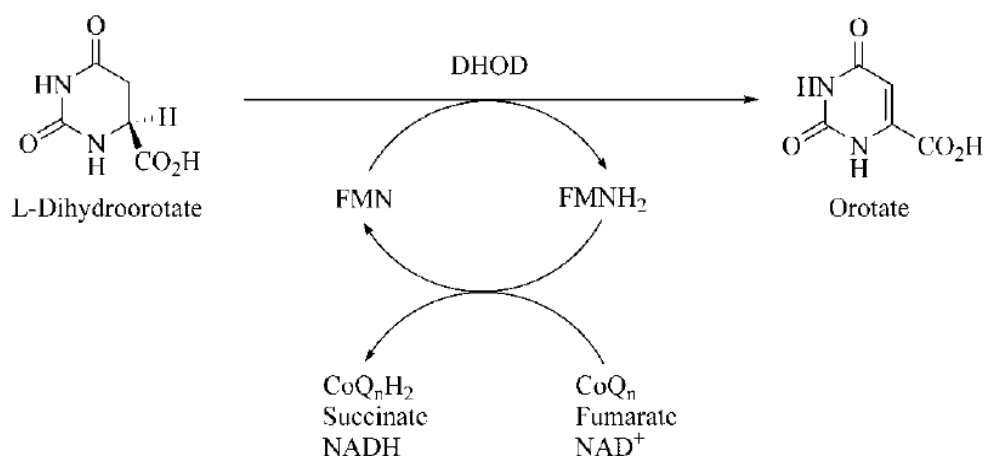


Figure 1.3: **Reaction catalyzed by DHOD.** L-Dihydroorotate is oxidized to orotate in the presence of flavin mononucleotide (FMN) co-factor, which is reduced to FMNH₂. FMNH₂ is oxidized by co-enzyme Q (CoQ or ubiquinone) to FMN which is re-used again in the reaction. CoQ itself is reduced to CoQH₂ (ubiquinol). The reaction yields two products, CoQH₂ and orotate. [Adapted from Patel *et al.* (2008)].

For HTS of small molecules active against the purified protein target, Patel and colleagues used the colorimetric reduction assay optimized for HTS (Baldwin *et al.*, 2005; Patel *et al.*, 2008). The assay measures the reduction of a chromogen, (DCIP), by CoQH₂ (Figure 1.3). DCIP is blue in its oxidized form and it turns colorless in its reduced form, thus enabling to measure the activity of DHOD as the production of orotate is equivalent to the production of CoQH₂ (Copeland *et al.*, 1995). In the presence of a small molecule that inhibits DHOD, there will be no color change in the reaction mixture, but a clear reaction mixture will result if a small molecule is unable to inhibit DHOD. The small molecules identified to be active against *Pf*DHOD through HTS were further analyzed using *in silico* structure-based docking for structural basis of inhibition (Patel *et al.*, 2008).

Availability of such protocols and assay information thus greatly facilitates the target and drug discovery processes. This data can be obtained from published literature and online public databases. ChEMBL (<https://www.ebi.ac.uk/chembl/db/>) is an example of a resource for assay information (Gaulton *et al.*, 2011). ChEMBL does not only contain assay data, but compound and target information is also available. The data in ChEMBL is manually curated

from published literature. BRENDA (<http://www.brenda-enzymes.org/>) is another source for assay information (Chang *et al.*, 2009). BRENDA contains a range of enzyme biological information, which includes reaction and specificity, isolation and preparation, enzyme structure, enzyme-disease relationships, functional parameters and organism related information. The data in BRENDA is curated from published literature and classified using the Enzyme Commission numbers (EC numbers).

1.3.3 Resistance

As mentioned before, resistance is a major problem in malaria. Most antimalarial drugs have failed due to the emergence of resistance. It is thus of great importance to take into account the possibility of drug resistance when assessing a potential target. Drug resistance can arise due to genetic mutations as seen with resistance to chloroquine, a very successful drug in the treatment of malaria (Djimé *et al.*, 2001; Sidhu *et al.*, 2002). Chloroquine acts by inhibiting the heme metabolism in the digestive vacuole of the parasite. The failure for the parasite to detoxify heme leads to heme buildup in the digestive vacuole which ultimately leads to death of the parasite. The study by Fidock *et al.* (2000) revealed that chloroquine resistance in *P. falciparum* was related to multiple mutations in 13-exon *pfert* gene located on chromosome 7, which encodes a digestive vacuole transmembrane protein *PfCRT* (Chloroquine Resistance Transporter).

The study was conducted *in vitro* using chloroquine-resistant and chloroquine-sensitive *P. falciparum* lines from Africa, South East Asia and South America. They identified eight different amino acid substitutions between the chloroquine-resistant and chloroquine-sensitive parasites at positions 74, 75, 76, 220, 271, 326, 356 and 371 (Figure 1.4). To determine the mutations that are crucial for chloroquine resistance, which can be used for monitoring, Djimé *et al.* (2001) analyzed the associations of the mutations in the *pfert* gene with chloroquine treatment in patients presenting with uncomplicated malaria in Mali. Their results showed that most of the patients with infections that persisted or reoccurred after treatment had parasites which harbored the *pfert* mutations at position 76, involving the substitution of Lysine with Threonine (K76T), suggesting that this mutation is an important marker for chloroquine resistance in *P. falciparum* malaria.

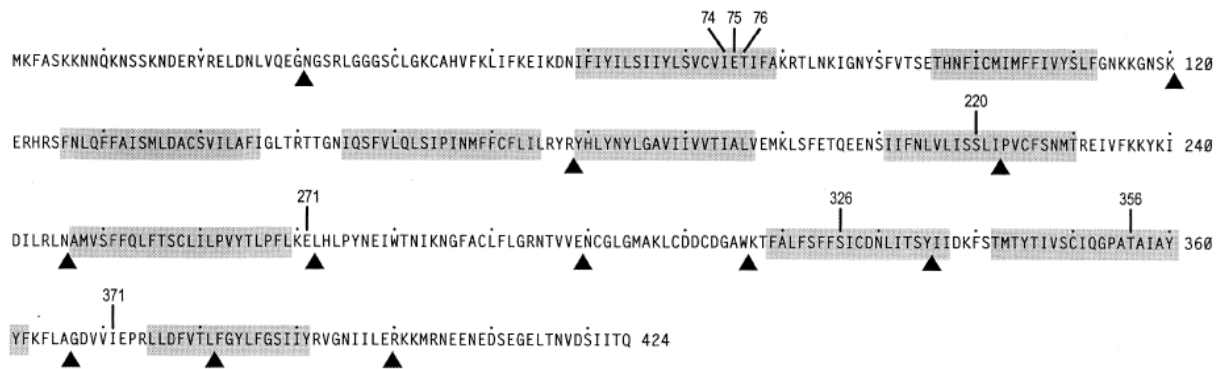


Figure 1.4: The 424 amino acid *PfCRT* transmembrane protein encoded by the 13-exon *pfert* gene. The shaded areas represent the predicted transmembrane segments, and the dark triangles represent the intron splice sites. The positions of the eight amino acid substitutions which distinguish between chloroquine-resistant and chloroquine resistant parasites are numbered. [Adapted from Fidock *et al.* (2000)].

Another mechanism by which resistance may arise is through the flexibility of metabolic pathways as mentioned in Section 1.3.1. If a selected target has isoforms (different forms of a protein), this may result in the substitution of the target with the isoform when the target is inhibited by a drug, which ultimately leads to resistance (Yeh *et al.*, 2004; Fatumo *et al.*, 2009). It is thus important to select targets that have no isoforms through careful analysis of literature and pathway information (choke-point analysis).

Isoforms in a species can also be identified through gene ontology (GO) annotations, a controlled vocabulary used to represent the biological aspects of a protein in an organism (Dimmer *et al.*, 2012). GO annotations provide descriptions for the “sub-cellular location”, “biological function” and “molecular function” of proteins. By analyzing gene products in a species that share the same GO terms at sub-cellular, molecular function and biological process levels, it is possible to identify isoforms that are involved in the same or similar reactions in metabolic pathways.

1.3.4 Toxicity

Selecting targets in a parasite causing a human disease that have homologs in human might cause toxicity. This could be caused by undesired binding of a drug to the homologous protein in human. It is thus important to investigate whether a selected target has a homologous protein in human and is unique to the parasite. Knowledge about orthology between human and parasite proteins can be obtained through OrthoMCL (Li *et al.*, 2003; Chen *et al.*, 2006). OrthoMCL is

a program that uses sequence similarity for grouping protein sequences into their orthologous groups. A web-based program (<http://www.orthomcl.org/>) is also available where users can view pre-computed ortholog groups or upload their own sequences for grouping (Fischer *et al.*, 2011).

However, orthology alone in determining toxicity is not sufficient enough to rule out a target with a homolog in human. For example, dihydrofolate reductase-thymidylate synthase (DHFR-TS) is present in both human and *Plasmodium* parasite, yet antifolates targeting DHFR-TS have been successful in the treatment of malaria (Zhang and Rathod, 2002). DHFR-TS is a bi-functional enzyme involved in the reduction of dihydrofolate (DHF) to tetrahydrofolate (THF). THF is an important co-factor in the biosynthesis of deoxythymidine monophosphate (dTMP), a precursor of DNA. In mammals, this enzyme is expressed as two separate proteins, DHFR and TS.

Zhang and Rathod (2002) associated antifolate selectivity to the differences in regulation and expression of DHFR-TS in human and parasite (Figure 1.5). In mammals, expression of DHFR and TS is hindered by the absence of substrate (or inhibitor), in which case the enzymes bind to their respective mRNA thus preventing translation. However, when the substrates (or inhibitors) are present, the enzymes dissociate from the coding regions of their mRNAs and translation is resumed. In *Plasmodium* parasites, however, this is not the case.

The inhibition of *Plasmodium* DHFR-TS and mRNA binding is not reversible as in mammalian DHFR and TS. The binding of antifolates to *Plasmodium* DHFR-TS does not cause the release of mRNA and thus the inactive enzyme cannot be replenished (Figure 1.5). The expla-

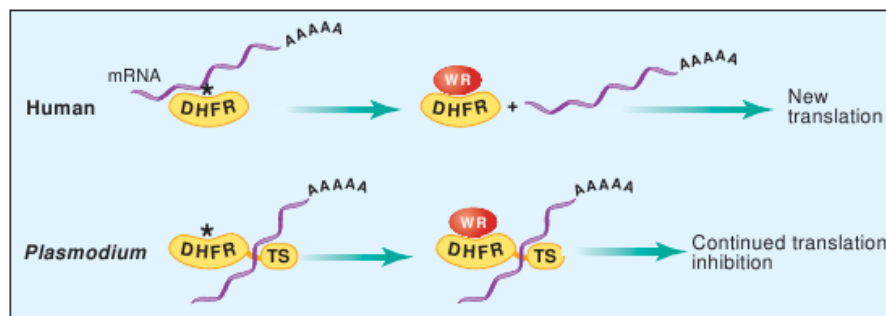


Figure 1.5: **Regulation and expression of human and *Plasmodium* DHFR.** The binding of inhibitors (red) on *Plasmodium* DHFR-TS does not cause the release of DHFR mRNA since mRNA binds elsewhere on the enzyme, unlike in the human DHFR where the mRNA is released on the binding of an inhibitor causing further protein synthesis. [Adapted from Goldberg (2002)].

nation for this phenomenon is that the binding of mRNA to *Plasmodium* DHFR-TS does not occur in the active site of the bi-functional enzyme, unlike DHFR and TS in mammals. These differences in binding sites and differences in regulation and expression of this target protein accounts for the selectivity of antifolates. It is thus important to analyze binding sites using crystal structures of homologs before a target can be ruled out in fear of unwanted binding of a drug to host proteins.

1.3.5 Structural information

Structural bioinformatics presents a wide range of computational techniques for drug discovery. These techniques require experimentally determined crystal structures of proteins for *in silico* analysis. The structures of proteins are identified through X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy techniques and are made available on the Protein Data Bank (PDB) database, accessible at <http://www.rcsb.org> (Berman *et al.*, 2000). In cases where experimentally determined structures of proteins are absent, homology and comparative modelling techniques are used for prediction. MODBASE (<http://salib.org/modbase>) is one such databases of modelled structures of proteins (Pieper *et al.*, 2009).

In silico docking programs are commonly used in structural bioinformatics to predict binding of small molecules to active sites on proteins; providing very useful information in drug discovery. The docking programs require crystal structures of proteins, preferably with a bound inhibitor in the active site of the resolved structure. Using the bound inhibitor, a range of small molecules are designed and tested *in silico* for their binding mode as well as binding affinity to target proteins. The presence of a bound inhibitor on crystal structures does not only aid in designing small molecule leads for drug discovery, but it also helps in determining the participating residues between the protein and ligand as well as confirming the binding modes (McGowan *et al.*, 2010).

Protein-ligand docking studies have been applied to the *P. falciparum* DHFR-TS enzyme, a malarial target in which mutations at residues 51 (N51I), 59 (C59R), 108 (S108N), and 164 (I164L) have made the parasite resistance to antifolates (Hunt *et al.*, 2005; Fogel *et al.*, 2008). In these studies, the crystal structure of the quadruple mutant DHFR with its bound inhibitor (WR99210) obtained from PDB (1J3K) was used. Small molecule ligands were designed based

on the structure of the WR99210 inhibitor and docked to mutant DHFR using the programs AutoDock (Hunt *et al.*, 2005) and GOLD (Fogel *et al.*, 2008) to identify and potential compounds that might be active.

The analogues of WR99210 were shown to interact with the mutant DHFR through hydrogen bonding with Asp54, Ile14 and Leu/Ile164 (Figure 1.6). Pro113 and Ile112 were also implicated to participate in the interactions, whereas mutations at Leu164 (mutated from Leu164) and Asn108 (mutated from Ser108) are said to be associated with the resistance to antifolates (Fogel *et al.*, 2008). Knowledge about the active site of the target helps in optimizing leads in drug discovery.

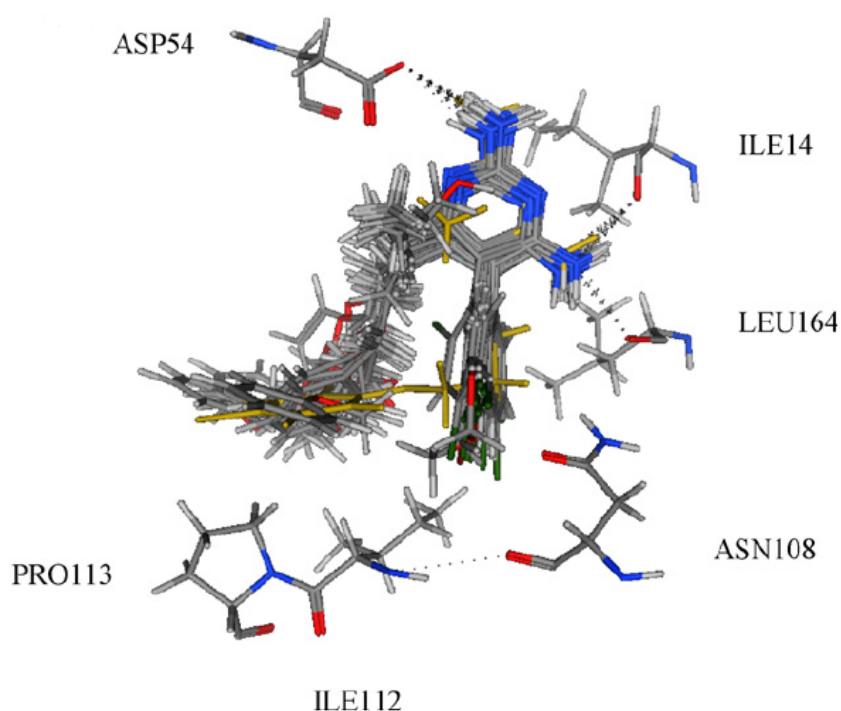


Figure 1.6: **Docking of WR99210 analogues to mutant DHFR.** The figure shows the docking of WR99210 (gold) and WR99210 analogues (in CPK) to the mutant DHFR. The compounds bind to the mutant DHFR through through hydrogen bonding with Asp54, Ile14 and Leu/Ile164. [Adapted from Fogel *et al.* (2008)].

More computationally intensive techniques have been developed for *in silico* docking of millions of known small molecules against *P. falciparum* targets (Kasam *et al.*, 2009). This sort of docking of compounds against targets is known as virtual HTS (vHTS). vHTS is used in combination with grid computing to minimize the time it would take to screen millions of compounds on a normal computer. The advantage to vHTS is that selection of hits is facilitated by the availability of chemical information since the small molecules being screened are known.

It is thus important to have an experimentally determined or modelled crystal structure of a protein when entering the target discovery process because *in silico* dockings to designed drugs can be performed in cases where *in vitro* and *in vivo* studies are not possible. *In silico* dockings are also useful when the drugs being tested are toxic to the model organism, where by computer modelling can be used to optimize the drug, to reduce toxicity, before testing in model organisms.

1.3.6 Druggability

Besides their utilization in docking studies, crystal structures of proteins are also utilized in the assessment of a target's "druggability". Druggability is defined as the ability of a protein to bind and be modulated by high affinity small molecules (Hajduk *et al.*, 2005; Coleman *et al.*, 2006; Cheng *et al.*, 2007). When assessing the druggability of protein, all binding sites are identified then assessed for their ability to bind small molecules with high affinity and specificity. A range of algorithms have been developed for identifying binding sites on 3D structures of proteins. These algorithms either use geometry (Hajduk *et al.*, 2005) or binding energy (Coleman *et al.*, 2006) to predict binding sites on the surface of proteins. Once the binding sites have been identified, they have to be assessed as being druggable or not druggable. This is the most challenging part in assessing druggability of a protein as there is no straightforward and cost effective way of doing this. Hajduk *et al.* (2005) used NMR-based screening data to derive characteristics that define druggable binding sites using 23 different proteins targets from different protein families. This was done by analyzing geometric parameters calculated for the binding sites identified by NMR-based screening of fragment library to derive an algorithm for predicting druggability.

In their study, the parameters that were analyzed for the prediction of the ability of pockets on proteins to bind to small molecules were total volume, polar and apolar surface area, total surface area, polar and apolar contact area, total contact area, roughness, the total number of charged residues, pocket compactness (defined as the ratio of pocket volume to pocket surface area) and principal moments which were for capturing the shape of the pocket (Hajduk *et al.*, 2005). They then compared the derived parameters of true positive (known ligand-binding sites on proteins identified by NMR) and negative pockets (algorithm-derived pockets with no

known binding ligand which were not identified by the NMR screening) where it was identified that there was no correlation between the individual parameters and hit rates (hit rates defined as the number of individual confirmed hits with K_D values less than 5mM divided by the total number of compounds screened as mixtures) observed in the NMR screens.

For this reason, linear and logarithmic regression dependency analysis on the parameters were performed in order to identify the relationships that correlate with the observed hit rates, and these were used to derive an algorithm for calculating the overall score of the pockets on their ability to bind small molecules. The algorithm (Algorithm 1.1) for calculating the druggability score was defined as a weighted linear combination of the linear and logarithmic dependencies on each of the pocket and protein binding site parameters;

$$score = \sum_{i=1}^N a_i X_i + b_i \log(X_i) \quad (1.1)$$

where N is the number of pocket and binding site parameters, X is the i th parameter and a_i and b_i are the weighted coefficients for the linear and logarithmic terms of the i th parameter, respectively (Hajduk *et al.*, 2005). The above model was tested on 35 protein targets, which had solved crystal structures and were known to bind high affinity molecules, for its ability to predict binding site druggability. The model was able to correctly predict 95% (33) of the 35 known ligand-binding sites to be druggable.

A different approach to druggability was used by Al-Lazikani *et al.* (2008) to assess druggability of proteins. In their method, binding sites were identified from crystal structures with bound ligands and predicted from protein structure analysis. An algorithm for calculating the physicochemical properties (volume, depth, curvature, accessibility, hydrophobic surface area and polar surface area) of the binding site was derived and trained on 400 proteins that bind small molecules that obey Lipinski's "Rule of Five" (a rule stating that orally active drugs have $MW < 500Da$, < 5 H-bond donors, < 10 H-bond acceptors and $\log P \leq 5$) (Al-Lazikani *et al.*, 2008). A decision tree was derived from training the algorithm, which predicts whether a binding site is druggable or not based on the calculated physicochemical properties.

This structure-based method of calculating druggability by Al-Lazikani *et al.* (2008) was further extended to calculate the druggability of all the crystal structures in PDB. The structures from PDB were classified into structural domains, then the binding sites in each of the identified

domains were predicted and their druggability calculated. The results for these druggability calculations are hosted in ChEMBL's DrugEBility website (<https://www.ebi.ac.uk/chembl/drugability/>). Users may search the database for predicted druggability of proteins using a UniProt accession, PDB code or by sequence similarity search as well as view the individual druggability of the domains and their predicted binding sites.

There are other methods however that do not rely on the presence of crystal structure for the prediction of druggability. Mapping of known proteins with known drug-like small molecules to whole genomes using orthology, a method based on the assumption that proteins having similar sequences are more likely to have similar conformation and thus bind the same small molecules, as well as using algorithms trained on a collection of known drug targets properties are ways of predicting druggability (Agüero *et al.*, 2008; Al-Lazikani *et al.*, 2008).

1.4 Problem statement

Current antimalarial methods have been effective in reducing malaria cases. However, the possible toxicity of repeatedly using artemisinin in humans and the emergence of artemisinin-resistant parasites creates a need to find new potential drug targets for the drug discovery process, in order to identify potential drugs to replace current drugs used against malaria. The process of drug discovery, however, is lengthy and sometimes fails due to the selection of drug targets that are not essential to the disease or drug targets which drugs cannot be designed for, that is proteins which are undruggable.

To limit the failure rate in malaria drug discovery, it is important to correctly identify and validate drug targets before they are entered into the drug discovery process through target discovery. As mentioned in Section 1.1, the conventional ways of target discovery, which are based on *in vivo* and *in vitro* techniques, can also take a long time and sometimes fail due to the inability to produce a viable organism or phenotype. For malaria, *in vivo* and *in vitro* target discovery is further hindered by the challenges related to genetic manipulation techniques for *Plasmodium* parasites and the difficulties of expressing proteins in *E. coli*. An alternative to these methods of target discovery is to identify and validate targets using an *in silico* approach.

This approach takes advantage of the large amounts of post-genomic data and resources available online. These resources use computational biology and bioinformatics techniques,

which are very useful tools for translating the amount of data available from sequencing genomes of many organisms into meaningful data that can be used to answer many biological questions that would sometimes take years to answer using traditional biological techniques. There are databases which are dedicated to assigning functions to proteins with unknown functions using sequence similarity algorithms. Similarity algorithms also help in identifying orthologs of proteins. Other databases are dedicated to pathway construction, and these pathways can be analyzed to identify the roles played by proteins in the disease and its progression.

Some databases are dedicated to curating and storing experimental data as well as data from literature. These data include assay information, protein isolation procedures, protein-protein interactions data, X-ray crystallography or NMR spectroscopy data as well as enzyme information. There are other databases that offer predictions of protein properties obtained via computational and bioinformatics techniques. These include protein-protein interaction predictions, domain function predictions, crystal structure predictions and druggability predictions. Proper utilization and mining of this data, together with experimental data from published literature could increase our knowledge on how the parasite interacts with its host and vector, as well as what makes it so successful during infection. Knowing how biological systems function could help us identify the most targets where attention should be focused in order to design drugs that are active against the malaria parasite.

1.5 Aims

The aim of this research was to extensively annotate the protein sequences from the *Plasmodium* parasites, *H. sapiens* and *A. gambiae* with as much data available from different online databases and resources which could aid in drug target selection in malaria. The annotation data was collected based on the six main target assessment criteria mentioned in Section 1.3, which are essentiality, assay feasibility, resistance, toxicity, structural information and druggability. Advanced data mining techniques were applied to store this data in a relational database and used to populate the current data in the Discovery system to allow for efficient filtering of protein sequences for the prioritization of malaria drug targets.

Chapter 2

Methods

2.1 Introduction

Discovery is a web-based system developed for the selection as well comparison of drug targets and lead compounds in malaria (Joubert *et al.*, 2009). The system provides a platform where researchers can view protein annotations and chemical compounds that are relevant in the selection of possible malaria drug targets and lead compounds, respectively, in a species-comparative manner. Users may search the database using either a protein or a chemical compound. Results from the searches are categorized into tabs where users can select and view the different types of data available. Although unique in that it provides available data from all three species involved in malaria, Discovery does have its pitfalls. This chapter describes the methods used in the rewrite and population of data in the new version of Discovery.

The current version of Discovery was developed using Python programming language and TurboGears web framework, with the data being stored in a MySQL database (Joubert *et al.*, 2009). The protein annotation data that is currently available in Discovery includes EC numbers, GO-terms, predicted domain functions, orthology, KEGG metabolism pathways, structural information and protein-protein interactions. To improve the performance of the new Discovery system as well as the user web interface, the programming language used was changed from Python to Java. However, the technical details of the programming and web page design will not be discussed here as it falls beyond the scope of this thesis. In the following sections of this chapter, the description of the sources of data used in populating Discovery will be explained followed by the methods used to integrate the data into Discovery.

2.2 Protein sequences and function

The ongoing genome sequencing projects release large amounts of raw genomic data. This data, however, gives the order in which the DNA in the chromosomes of the organisms is organized. On its own, it does not provide any insights into the organism being sequenced. The data has to be translated into meaningful information by determining the regions which code for proteins, and the proteins themselves have to be assigned function. There are functional genomic databases which work together with genome sequencing projects to annotate the data as it is being produced for different organisms. This data is stored and made available to the public and is continually updated as new sequencing data is produced. PlasmoDB (<http://plasmodb.org/>) is one such database dedicated to the annotation of the *Plasmodium* species (Bahl *et al.*, 2002; Aurrecoechea *et al.*, 2009).

PlasmoDB

PlasmoDB is the official database for the *P. falciparum* genome sequencing consortium (Bahl *et al.*, 2002). It is one of the many genome resources hosted by the EuPathDB (<http://eupathdb.org/>) and contains fully sequenced genomes of *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi* and *P. knowlesi* as well as other incomplete genomes of *P. reichenowi* and *P. gallinaceum* (Aurrecoechea *et al.*, 2009). The PlasmoDB database was designed to store complete and incomplete genomic sequences produced by the *Plasmodium* genome projects together with analysis, annotations and post-genomic data available for the parasite. Transcript expression data, protein expression, putative protein function, protein interactions, population biology, evolutionary data, protein features and protein localization are datasets available in PlasmoDB are obtained from different independent resources. A user-friendly web interface allows users to access the data in PlasmoDB and also allows users to download desired data. Users can either search the database using a gene ID or a keyword to retrieve data available for that particular gene. Users may also combine and filter searches allowing them to answer specific questions.

VectorBase

Another genome information system, VectorBase (<http://www.vectorbase.org>), is available for the annotation of invertebrate vector genomes that are responsible for the transmission of disease causing pathogens to humans (Lawson *et al.*, 2009). The vector genomes that are currently available in VectorBase include *A. gambiae*, *Aedes aegypti*, *Culex quinquefasciatus*, *Ixodes scapularis*, *Pediculus humanus*, *Rhodnius prolixus*, *Glossina morsitans morsitans*, *Lutzomyia longipalpis* and *Phlebotomus papatasi* (Lawson *et al.*, 2009). Of all these species, *A. gambiae* is relevant to this research as it is responsible for the transmission of the malaria causing parasite to humans (Holt *et al.*, 2002). VectorBase works together with a number of genome sequencing centers in early annotation of new vector genome sequences as well as re-annotation using manual and automatic approaches.

A community annotation pipeline (CAP) has been developed in VectorBase in order to improve the quality of the annotations. In this system, representatives with biological knowledge on the species and informatics skills work with the community to curate and increase the quality of the annotation data submitted. Manual annotations from within VectorBase and those submitted by the community are stored in a Chado database and are subjected to quality and consistency checks by the community representatives to ensure that the data going into VectorBase is corrected for errors and redundancy (Lawson *et al.*, 2009). VectorBase also provides a tool for the comparison of the genomes of the three species of mosquitoes (*A. gambiae*, *A. aegypti* and *C. quinquefasciatus*), gene expression data as well as ontologies describing the mosquito and tick anatomies.

Ensembl

A more comprehensive genome information system, Ensembl (<http://www.ensembl.org>), is available for the storage, integration analysis and visualization of chordate genomes (Hubbard *et al.*, 2009). This system focuses on vertebrate genomes, mostly mammalian, as well as genomes from selected model organisms and disease vectors. The Ensembl system provides annotated gene sets for genomes generated automatically through a pipeline as well as comparative genomics data (sequence alignments, ortholog assignment, paralog assignment, gene trees) between genes and genomes, also generated by an automatic pipeline

(Hubbard *et al.*, 2009). The data in Ensembl is open to the public and can be accessed via the web interface, downloaded through BioMart or the file transfer protocol (FTP) site (<http://www.ensembl.org/info/data/ftp/>) and through application programming interfaces (API's).

UniProt

The efforts of the different genome information systems to translate raw genomic data into predicted protein sequences and subsequent annotations offers a large amount of data to the biological community. Storing, integration and standardization of the protein sequences and annotation data from these different resources is important to biological research. The Universal Protein Resource (UniProt, <http://www.uniprot.org>) is a comprehensive database for protein sequence and functional annotation (Consortium, 2012). There are four components making up UniProt; the UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef), UniProt archive (UniParc) and UniProt Metagenomic and Environmental Sequences database (UniMES)(Consortium, 2008, 2012). UniProtKB is the main component of UniProt and contains two sections i.e., UniProtKB/Swiss-Prot (records with manual and non-redundant annotations) and UniProtKB/TrEMBL (records with computationally generated unreviewed annotations) (Magrane and Consortium, 2011).

Protein annotations in UniProtKB include protein names, taxonomy, function, catalytic activity, pathways, GO annotations as well as sequence features (Consortium, 2008; Magrane and Consortium, 2011). UniProtKB also provides extensive cross-referencing to other databases, including organism-specific databases, structural databases and disease databases. UniRef contains three datasets of closely related sequences merged according to 100% (UniRef100), 90% (UniRef90) and 50% (UniRef50) sequence identity (Consortium, 2008, 2012). These three datasets are used to reduce sequence redundancy, thereby increasing the speed of sequence similarity searches and reduce bias. UniParc is the main repository and storehouse of protein sequences from different sources. UniParc ensures that there is no redundancy in the sequences submitted to UniProt and also keeps the history of all protein sequences. The UniMES component of UniProt provides metagenomic data.

The UniProt web interface offers users with a variety of tools to interact with the data avail-

able. Information about a protein can be retrieved via a text-based query or sequence similarity search; the results are configurable and can also be downloaded (Magrane and Consortium, 2011). Multiple sequence alignment and batch retrieval tools are also available. Another useful tool offered in UniProt is the identifier mapping tool, which is used to map UniProt identifiers to other identifiers used in the cross-referenced databases and vice versa. Different data types and in different formats are freely available in UniProt and can be downloaded via the UniProt FTP server (<ftp://ftp.uniprot.org/pub>) (Consortium, 2012).

UniProt-GOA

The GO annotations provided in UniProt are provided by the Gene Ontology Annotation project (UniProt-GOA) hosted at the European Bioinformatics Institute (EBI) (Dimmer *et al.*, 2012). This project is a result of a collaboration between the UniProtKB and the Gene Ontology project (<http://www.geneontology.org>, (Ashburner *et al.*, 2000)). UniProt-GOA provides evidence-based manual and automated associations of GO terms (which describe the sub-cellular location, biological processes and molecular functions of proteins) from the Gene Ontology database with the proteins in UniProtKB. These associations can be viewed via the QuickGO tool (<http://www.ebi.ac.uk/QuickGO>) or downloaded via the EBI FTP server (<ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/>) (Dimmer *et al.*, 2012).

InterPro

Determining the function of a protein experimentally can be quite laborious and it is sometimes not possible due to the inability to recombinantly express certain proteins. To overcome this problem, databases that assign function to proteins using pattern recognition methods have been designed. These pattern recognition or signature databases use different methods and have different biological focus. Individually searching of these databases for predicted functions and properties of particular protein can be time consuming and laborious. A single database, InterPro (<http://www.ebi.ac.uk/interpro/>), was designed to integrate these signature databases in order to determine information about the protein families, domain and functional sites (Quevillon *et al.*, 2005; Hunter *et al.*, 2009).

Curators in InterPro group signatures that describe the same protein family, domain or

functional site into a single InterPro identifier (Apweiler *et al.*, 2001). The InterProScan tool was designed to integrate all the signature recognition methods in InterPro into a single application which users can use to find signatures that match their protein of interest in order to assign function. Users can use the web-based version of InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>) or the stand-alone version available for download at the EBI FTP server (<ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan/>). Signature databases and tools currently integrated in InterPro include Pfam, PRINTS, PROSITE, SMART, ProDom, PIRSF, SUPERFAMILY, PANTHER, CATH-Gene3D, TIGRFAMs, HAMAP (Quevillon *et al.*, 2005; Hunter *et al.*, 2012).

2.2.1 Obtaining protein sequences

The proteome sequences for *P. berghei*, *P. chabaudi*, *P. falciparum*, *P. knowlesi*, *P. vivax* and *P. yoelii* were downloaded from PlasmoDB (<http://plasmodb.org/common/downloads/>) in FASTA formats. The *H. sapiens* and *A. gambiae* proteome sequences were downloaded from Ensembl (ftp://ftp.ensembl.org/pub/release-66/fasta/homo_sapiens/pep/) and VectorBase (ftp://ftp.vectorbase.org/public_data/organism_data/agambiae) FTP sites, respectively, also in a FASTA format. A protein identifier mapping file containing mappings of the UniProt accessions to protein identifiers used in other databases was downloaded from UniProtKB FTP server (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping). The file contains three columns separated by tabs; the first column contains the UniProt accession, the second column contains the cross-referenced database and the last column contains the protein identifier in the cross-referenced database. The protein identifier mapping file was used to assign UniProtKB accessions to the human, parasite and mosquito protein sequences in Discovery. In cases where there was more than one UniProt accession for a particular protein, no UniProt accession was assigned. The proteins were linked to their original databases using their protein identifiers and also to UniProt where UniProt accessions existed.

2.2.2 Functional annotation

Prediction of functional motifs and features of the protein sequences was carried out using a stand-alone version of InterProScan (v4.8). Protein sequences in FASTA format were used as input for the InterProScan program and the output format was set to “raw” to allow the results to be integrated into the database. The species-specific GO annotations files for *P. falciparum* (493.P_falciparum.goa), *P. knowlesi* (31342.P_knowlesi.goa), *P. vivax* (31632.P_vivax.goa), *P. yoelii* (21631.P_yoelii.goa), *H. sapiens* (25.H_sapiens.goa) and *A. gambiae* (22426.A_gambiae.goa) proteomes were downloaded from the UniProt-GOA FTP site (<ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/>). The GO annotation files were in GAF 2.0 file format. UniProt accession were used to assign the GO annotations to the protein sequences in the database. The AmiGO visualization tool (Carbon *et al.*, 2009) was used to visualize the GO terms in graph form. Links to the Gene Ontology database were created for the GO terms.

2.3 Orthology

OrthoMCL

Identification of homologous (related) proteins sequences between and within species provides valuable information for genome annotation. Homologous protein sequences can be orthologs (result from speciation events between species) or paralogs (results from gene duplication within a species). Paralogs may share sequence similarity, however their functions may not be retained in the duplicated gene. Orthologs on the other hand share sequence similarity as well as function, and this makes identification of orthologous groups between species important in genome annotation as it allows proteins function to be inferred from other proteins with known function .

The OrthoMCL database (<http://www.orthomcl.org>) was created to store predicted orthologous groups from different organisms (Li *et al.*, 2003; Chen *et al.*, 2006). The predicted clusters of orthologous groups in the OrthoMCL databases are calculated using the OrthoMCL algorithm (Li *et al.*, 2003). The algorithm (Figure 2.1) clusters orthologs by first running a all-against-all BLAST on the protein sequences of all the species to be compared. It then identifies

possible ortholog pairs as reciprocal best hits across two genomes; in other words, two proteins assumed to be orthologs if they both find each other as best hits in the opposite genomes by BLAST analysis.

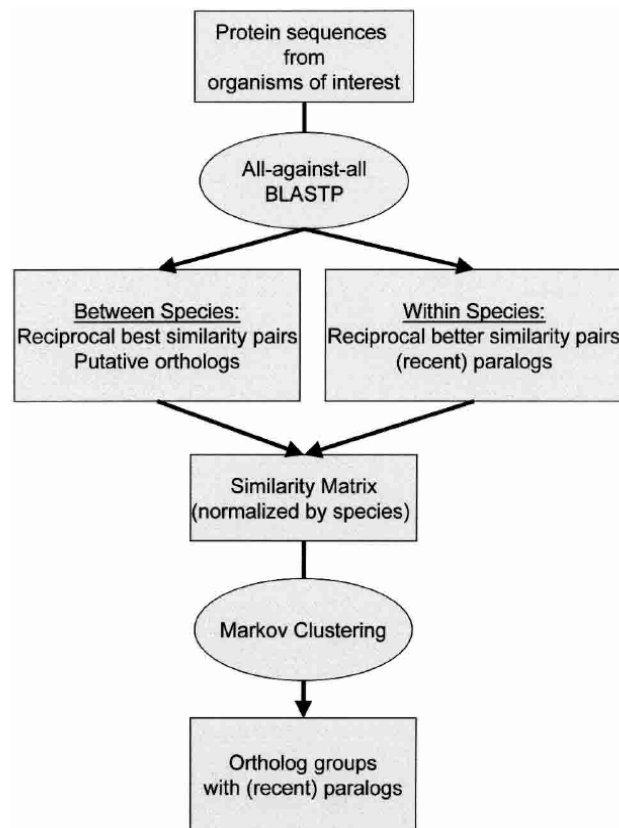


Figure 2.1: **Clustering of orthologs from different organisms using the OrthoMCL algorithm.** [Adapted from Li *et al.* (2003)].

The algorithm then identifies possible paralogs within each genome for each identified ortholog as protein sequences that are more similar to each other than they are to any other sequence from the other genomes (Li *et al.*, 2003; Chen *et al.*, 2006). The predicted orthologous and paralogous relationships are then converted into a graph, where protein sequences are represented by nodes and the relationships are represented by weighted edges. The weights of the edges in the graph are then normalized to reduce bias, and the graph converted to a similarity matrix. A Markov Clustering algorithm (MCL) is then applied to the similarity matrix to create orthologous groups (Fig 2.1).

Currently, the OrthoMCL database (version 5) contains 150 genomes, 1 398 546 protein sequences and 124 740 ortholog groups (<http://www.orthomcl.org>). Users can search the OrthoMCL database for a protein and its predicted orthologous groups using protein identifier,

keywords or by BLAST search. An option to upload a set of protein sequences for mapping to OrthoMCL groups is also available. Alternatively, a stand-alone OrthoMCL software can be downloaded and installed locally to assign proteins sequences from genomes of interest into ortholog groups.

2.3.1 Assignment of sequences to orthologous groups using OrthoMCL

To identify the groups of orthologous proteins in the Discovery database, a stand-alone version of OrthoMCL (v2.0.2) was used. The proteome sequences of each of the species (in separate FASTA files) were used as input for the command-line-based OrthoMCL program. The resulting file (groups.txt) contained a group of orthologous proteins in each line, with the proteins represented by their protein identifier and not the complete protein sequence.

2.3.2 Multiple sequence alignment using T-coffee

To identify the conserved and variant sites in the protein sequences within the ortholog groups generated by OrthoMCL, a multiple sequence alignment was done using T-coffee (Version 8.99). T-coffee (Tree-based Consistency Objective Function for alignment Evaluation) is multiple sequence alignment program that uses a heuristic search to generate an alignment (Notredame *et al.*, 2000). FASTA files containing the protein sequences belonging to each orthologous groups (generated by OrthoMCL) were compiled and individually aligned using T-coffee (default parameters used) to produce a multiple sequence alignments. InfoAlign (Rice *et al.*, 2000) was used to obtain the information about the multiple sequence alignment. The Jalview applet was used to view the multiple sequence alignments.

2.4 Structural information

PDB

The PDB (<http://www.rcsb.org>) is a resource dedicated to the deposition of crystal structures of biological macro-molecules which was established at Brookhaven National Laboratories (BNL) (Berman *et al.*, 2000). The Research Collaboratory for Structural Bioinformatics (RSCB) is responsible for handling and managing the data deposited in PDB. The crystal struc-

tures submitted in PDB are obtained from techniques such as NMR, cryoelectron microscopy, X-ray crystallography and theoretical modelling. These structures (atomic coordinates, factors and method of determination) submitted to PDB have to be assessed for quality before made available to the public to increase the integrity of the data in PDB. This is an active process involving both the PDB curators and the author of the submitted structure.

The submitted structure is assigned a PDB identifier and then annotated and validated according to the PDB standards. The structures are checked for standard covalent bonds distances and angles. The correct stereochemistry of proteins (or nucleic acids) and nomenclature for ligands and atoms is checked. A protein sequence is derived from the coordinates of the structure and checked for accuracy; any inconsistencies or redundancies that occur between the derived sequence and other sequences in the database are resolved (Berman *et al.*, 2000). Throughout data processing, communications with the author and changes made to the structure are recorded.

Users can browse or search (via chemical or structure) data available in PDB through the web-interface (<http://www.rcsb.org>). The results for searches can be sorted or filtered according to the users preference to allow handling of large datasets (Rose *et al.*, 2011). The structure records in PDB are can be visualized using the Jmol viewer (<http://www.jmol.org>) and are also linked to literature references. PDB offers sequence and structure analysis tools as well as pre-calculated alignments for representative protein chains (Rose *et al.*, 2011). Widgets and application programming interface (API) services are also offered by PDB for web developers to enable access to the data.

MODBASE

For protein sequences with unknown crystal structures, homology or comparative modelling can be applied to the sequences to predict their crystal structures. MODBASE (<http://salilab.org/modbase>) is one such database dedicated to the modeling of crystal structures of proteins which do not have experimentally determined crystal structures (Pieper *et al.*, 2009). The predicted crystal structures of proteins in MODBASE are calculated using an automatic pipeline software called MODPIPE. The MODPIPE software pipeline is dependent on a number of modules available from the MODELLER program for the calculation of crystal structures of

proteins (Pieper *et al.*, 2009).

Comparative modelling relies on the availability of a crystal structure that is similar/related to the target protein for which a structure is to be predicted. The MODELLER program finds template structures (from structure databases) similar to the target protein, which are then used to generate sequence-structure alignments. From each of the sequence-structure alignments, models are built and a representative model for each of the alignments is chosen based on the Discrete Optimized Protein Energy (DOPE), a statistical potential based on atomic distance (Pieper *et al.*, 2009). The reliability of the models is then evaluated based on the coverage of the modelled sequence, sequence identity (sequence-structure alignment), gaps in the sequence-structure alignment, compactness of the model and different statistical *Z-scores* (Pieper *et al.*, 2009). The predicted MODBASE structures can be accessed via the web interface available at <http://salilab.org/modbase> by querying with gene names, gene identifiers or PDB identifiers. Selected model predictions for whole genomes are also available for download via the FTP site (<ftp://salilab.org/databases/modbase/projects>).

2.4.1 BLAST search against PDB database

A similarity search using BLAST (version 2.2.25) was run against PDB proteins to identify crystal structures of the proteins in the Discovery database and those that are similar to them. An *E*-value cut-off value of $1e^{-6}$ and minimum sequence coverage of 70% were used. The Jmol applet was used to view the crystal structures from PDB.

2.4.2 Predicted MODBASE structures

The predicted model structures for *P. falciparum*, *H. sapiens* and *P. vivax* were downloaded from the MODBASE FTP site (<ftp://salilab.org/databases/modbase/projects>) along with the summary files. The Jmol applet was also used to view the MODBASE models. The summary files were used to display the details of the predicted model structures. For more details on the predicted structures of proteins, links to the models in the MODBASE database were created using the UniProt accessions.

2.5 Metabolic pathways and enzyme information

KEGG

The KEGG database (<http://www.genome.jp/kegg/>) is an integrated resource designed for the assignment of genes from completely sequenced genomes to higher-level systemic functions of the cell, organism and the environment (Kanehisa and Goto, 2000). This information is collected from the genomic and molecular information available from researches and used to understand the biological systems. KEGG consists of different databases used to organize data. The genes from all the completely sequenced genomes, along with the gene information, are stored under the KEGG GENES database (Kanehisa *et al.*, 2012). The genes are linked to the pathway maps available under the KEGG PATHWAYS database, which contains a collection of pathway maps for metabolism, signal transduction, genetic information processing, environmental information processing, organismal systems, human diseases, drug development and cellular processes (Kanehisa *et al.*, 2006, 2012).

Reference pathways in KEGG PATHWAYS are manually drawn to represent molecular interactions and reactions. The reference pathways are a network of enzymes (represented by EC numbers) derived from well studied organisms and are used to computationally generate organism-specific pathways through sequence similarity (Kanehisa and Goto, 2000). The EC numbers in the pathway maps are linked to enzyme information stored under the KEGG ENZYME database. KEGG contain 15 main databases (Table 2.1) that are categorized into systemic, genomic and chemical information (Kanehisa *et al.*, 2012). The different databases in KEGG are the basis of organization of data and storage, and each database is focused yet related to other databases.

MPMP

Although KEGG is a comprehensive resource for metabolic pathways, it lacks some information that is specific to the metabolic pathways of the *Plasmodium* parasites. The MPMP database (<http://sites.huji.ac.il/malaria/>) was constructed to incorporate metabolic pathway maps and information relevant to the malaria parasites (Ginsburg, 2006, 2009). The metabolic pathways in MPMP were manually constructed using the classical biochemical path-

Table 2.1: Different databases integrated in KEGG.

Category	Database	Content
Systems information	KEGG PATHWAY	Pathway maps
	KEGG BRITE	Functional hierarchies
	KEGG MODULE	KEGG modules
	KEGG DISEASE	Human disease
	KEGG DRUG	Drugs
	KEGG ENVIRON	Crude drugs
Genomic information	KEGG ORTHOLOGY	KO groups
	KEGG GENOME	KEGG organisms
	KEGG GENES	Genes in high quality genomes
Chemical information	KEGG COMPOUNDS	Metabolites and other small molecules
	KEGG GLYCAN	Glycans
	KEGG REACTION	Biochemical reactions
	KEGG RPAIR	Reactant pairs
	KEGG RCLASS	Reaction class
	KEGG ENZYME	Enzyme nomenclature

ways from KEGG as templates. Irrelevant information from the pathways from KEGG were removed and some pathways were combined into more comprehensive and interconnected pathways. The metabolic pathways were further enriched with information obtained from other *Plasmodium* specific databases, like PlasmoDB, and displayed in a user friendly and informative way (Figure 2.2). The enzymes in the metabolic pathways are displayed with color-coded 48 hour transcriptomic clocks, which show stage-dependent over-transcription (red) or under-transcription (green) of a gene coding for the enzyme (Figure 2.2).

Additional links to other relevant databases are provided in the metabolic pathways. The enzymes in each pathway are linked to enzyme databases BRENDA (<http://www.brenda-enzymes.org/>), ExPASy (<http://enzyme.expasy.org/>) and the reaction schemes at the International Union of Biochemistry and Molecular Biology (IUBMB, <http://www.chem.qmul.ac.uk/iubmb/enzyme/reaction/>) for further information (Ginsburg, 2006). There are also links to the genome databases PlasmoDB and GeneDB (<http://www.genedb.org/>). The transcriptome clock is linked to the DeRiSi Lab Malaria Transcriptome database (<http://malaria.ucsf.edu/>). The MPMP database does not only contain metabolic pathways; transport functions, cell-cell interactions, protein trafficking, morphological development of

Nitrogen metabolism

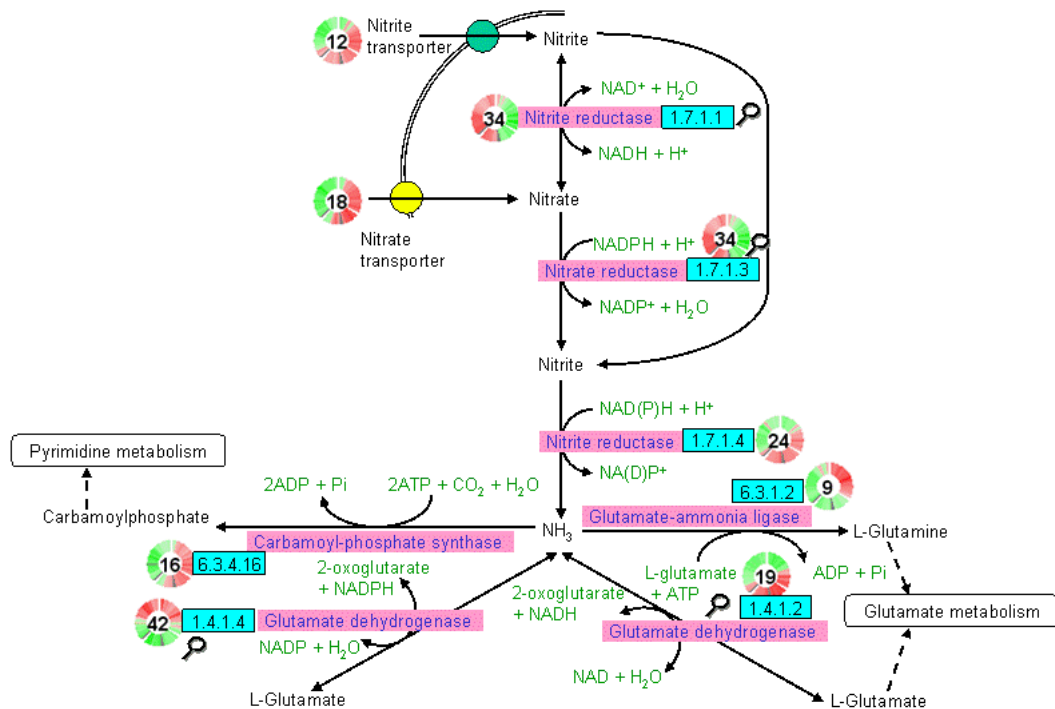


Figure 2.2: **Nitrogen metabolism pathway for the *Plasmodium* parasite.** An example of a metabolic map from MPMP showing enzyme names (in pink background), EC numbers (in blue background) and transcriptomic clocks. EC numbers are linked to external enzyme databases whilst transcriptomic clocks are linked to the transcriptome database for malaria (Adapted from <http://sites.huji.ac.il/malaria/maps/nitrogenmetpath.html>).

blood forms, as well as invasion and motility are also included (Ginsburg, 2006). Users may browse the metabolic pathways in the database or search by enzyme names, compounds, protein names, EC number or *Plasmodium* identifiers.

Reactome

Apart from KEGG, which focuses on pathways of all fully sequenced genomes, and MPMP, which focuses only on the pathways of the malaria parasites, Reactome (<http://www.reactome.org>) is another pathway database available for manually curated and peer-reviewed human pathways (Joshi-Tope *et al.*, 2005). What makes Reactome unique is its underlying data model used to represent different processes of the human system (Joshi-Tope *et al.*, 2005;

Vastrik *et al.*, 2007). The reactions in Reactome are treated as basic units, with each reaction linked to its appropriate literature for evidence. These reactions are grouped to form pathways based on relationships and inter-dependency amongst each other i.e. pathways may contain sequential, parallel or cyclic reactions. This sort of data organization found in Reactome allows for interconnection and representation of various biological processes in the human system including metabolic pathways, regulatory pathways, signal transduction and cell cycle (Joshi-Tope *et al.*, 2005; Vastrik *et al.*, 2007).

However, Reactome does not only contain human reactions and pathways. Using protein sequence similarity derived from OrthoMCL, human pathways in Reactome are used to infer pathways of other selected species based on orthology (Joshi-Tope *et al.*, 2005; Vastrik *et al.*, 2007). *P. falciparum* is amongst the species in Reactome with reactions and pathways inferred through sequence similarity. The data from Reactome is freely available and users can search or browse pathways through the user interface available at <http://www.reactome.org> and can also use different pathway, expression and comparative analysis tools available (Haw *et al.*, 2011). Different datasets of Reactome and code can be downloaded through the download site (<http://www.reactome.org/download/>).

ExPASy

ExPASy (Expert Protein Analysis System), available at <http://www.expasy.org/>, is a web portal providing databases and tools for the analysis of proteins and proteomics (Gasteiger *et al.*, 2003). The web portal is hosted by the Swiss Institute of Bioinformatics (SIB). The ENZYME database (<http://enzyme.expasy.org>), which provides information related to enzyme nomenclature, is amongst databases hosted by ExPASy (Bairoch, 2000). The enzyme records in the ENZYME database are provided with an EC number along with a recommended name, alternative names, catalytic activity, co-factors as well as cross-references to other relevant databases. The data in the ENZYME database can be accessed through the web interface or downloaded through the FTP server (<ftp://ftp.expasy.org/databases/enzyme/>).

BRENDA

The Braunschweig Enzyme database (BRENDA) is comprehensive system providing data for all enzymes classified according to EC number scheme (Schomburg *et al.*, 2002; Barthelmes *et al.*, 2007). The enzyme records in BRENDA are manually curated from literature and made available at <http://www.brenda-enzymes.org>, along with tools for querying and analyzing enzymatic data. BRENDA also links each enzyme records to its literature references as well as the biological source in which the enzyme was extracted from (organism, organ, tissue or cellular localization). BRENDA covers all molecular and biochemical aspects of enzymes including pathways, classification and nomenclature, reaction and specificity, functional parameters, organism information, enzyme structure, isolation and preparation information, literature references as well as enzyme-disease information (Barthelmes *et al.*, 2007; Scheer *et al.*, 2011).

2.5.1 Metabolic pathway assignment

Pathways in which the proteins in the Discovery database are involved in were assigned by creating links to the pathway databases. For pathways in the KEGG database, the tab-delimited files for *A. gambiae* (aga_gene_map.tab), *H. sapiens* (hsa_gene_map.tab), *P. berghei* (pbe_gene_map.tab), *P. chabaudi* (pcb_gene_map.tab), *P. falciparum* (pfa_gene_map.tab), *P. knowlesi* (pkn_gene_map.tab), *P. vivax* (pvx_gene_map.tab) and *P. yoelii* (pyo_gene_map.tab), containing all the genes for each species and the pathway maps each gene is involved in were downloaded from the KEGG FTP site. These tab-delimited files for each of the species were obtained from KEGG when they were still publicly available.

Links to the organism specific pathways were constructed using the following link construction: http://www.genome.jp/kegg-bin/show_pathway?<PATHWAY>+<KEGG_ID>, where <PATHWAY> is a combination of three letter prefix for an organism and a five digit number for a pathway, and <KEGG_ID> is the protein identifier used in KEGG. For example, a link to the pathway map “00260” for the *P. falciparum* protein “MAL13P1.67” would be: http://www.genome.jp/kegg-bin/show_pathway?pfa00260+MAL13P1.67.

The PlasmoDB page for each of the *P. falciparum* proteins contains pathway links to the MPMP database. These links were extracted programmatically and used to assign pathways to the *P. falciparum* proteins in the Discovery database. For Reactome, path-

ways for which the proteins are involved in were searched using the following link construction: <http://www.reactome.org/cgi-bin/link?SOURCE=UniProt&ID=<UniProtAC>>, where UniProtAC is the UniProt accession number for the protein. If pathways and reactions for that particular UniProt accession exist, the links to the pathways and reactions were extracted and assigned to the protein.

2.5.2 EC number assignment linking to databases

An ENZYME database data file (enzyme.dat) was downloaded from the ExPASy FTP server and used to assign EC numbers to the proteins in the Discovery database. For the *Plasmodium* species, however, the EC numbers were obtained programmatically through the PlasmoDB page for each parasite proteins where it existed. Once the EC numbers were assigned to the proteins, links to external enzyme databases (KEGG ENZYME and BRENDA) were created. For links to the BRENDA database, the following link construction was used: [http://www.brenda-enzymes.info/php/result_flat.php4?ecno=<ECNUMBER>&Suchword=&organism\[\]=<ORGANISM>&show_tm=0](http://www.brenda-enzymes.info/php/result_flat.php4?ecno=<ECNUMBER>&Suchword=&organism[]=<ORGANISM>&show_tm=0), where <ECNUMBER> is the EC number for a particular protein and <ORGANISM> can be “Homo+sapiens”, “Anopheles+gambiae”, “Plasmodium+berghei”, “Plasmodium+chabaudi”, “Plasmodium+falciparum”, “Plasmodium+knowlesi”, “Plasmodium+vivax” or “Plasmodium+yoelii”. Links to KEGG ENZYME were constructed as follows: http://www.genome.jp/dbget-bin/www_bget?ec:<ECNUMBER>, where <ECNUMBER> is also the EC number for a protein.

2.6 Protein-protein interactions

Understanding the molecular interactions within cells is important in revealing information on the biological processes that occur. Most molecular interactions within a cell occur between proteins. Determination and understanding of these protein-protein interactions does not only give insights into biological processes that occur within a cell, but it also helps in understanding the biological functions of proteins. A number of databases have been developed to identify and store such protein-protein interactions. The Database of Interacting Proteins (DIP, <http://dip.doe-mbi.ucla.edu>), the Molecular Interactions database (MINT, <http://mint.bio.uniroma2.it/mint/>) and IntAct (<http://www.ebi.ac.uk/intact/>) are

examples of databases that focus on the identification, curation and storing of protein-protein interactions from published peer-reviewed journals (Salwinski *et al.*, 2004; Ceol *et al.*, 2010; Kerrien *et al.*, 2012).

These databases form part of the International Molecular Interaction Exchange consortium (IMEx, <http://www.imexconsortium.org/>) which focuses on reducing overlap/duplication of interaction data by the sharing of data to be curated amongst the participating databases as well as exchanging completed interaction data. The interaction data released by these databases is compliant with the Proteomics Standard Initiative-Molecular Interaction (PSI-MI) standard syntax and semantics for data representation and is available for download via the FTP servers of the participating databases.

2.6.1 Assignment of protein-protein interactions

The files containing the binary interactions were downloaded from the interaction databases DIP (<http://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=3>), MINT (<ftp://mint.bio.uniroma2.it/pub/release/mitab26/current/2010-12-15-mint-full-binary.mitab26.txt>) and IntAct (<ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact.zip>). The UniProt accessions were used to assign the interaction data to the proteins in Discovery. For every interaction, the columns for both the interactors, the detection method used, interaction type, taxonomy for both interactors, database identifier and interaction scores were selected. The database identifiers were used to link the each binary interaction to its original database.

2.7 Druggability

DrugEBIity

The ChEMBL database (<https://www.ebi.ac.uk/chembl/db/>), hosted by the EBI, is a comprehensive database containing information for bioactive molecules with drug-like properties (Gaulton *et al.*, 2011). The information in ChEMBL is manually curated from peer-reviewed literature. The data available in ChEMBL includes compound structures, assay information, activity information and targets information. ChEMBL also hosts the DrugEBIity database

(<https://www.ebi.ac.uk/chembl/drugability/structure>); a structure-based druggability search engine. Users can upload their 3D structure of a protein on the DrugEBility database to calculate its predicted druggability. They can also search pre-calculated druggability scores using domain numbers, UniProt accessions or by BLAST search.

2.7.1 BLAST search against DrugEBility database

The complete sets of proteins sequence for the genes and domains from the DrugEBility database were kindly provided by the ChEMBL group. Using the protein sequences in Discovery, a local BLAST search against the DrugEBility protein was done, first against the gene sequences and then with the domain sequences using an E -value cut-off of $1e^{-6}$ and a minimum sequence coverage of 70%. For each of the BLAST hits (gene and domain sequences), a link to the DrugEBility data base was created to allow the user to view detailed druggability calculations for each hit. Links for the gene hits were created as follows: https://www.ebi.ac.uk/chembl/drugability/protein/<UNIPROT_AC>, where UNIPROT_AC is the UniProt accession for the hit. The links for the domain hits were created as follows: <https://www.ebi.ac.uk/chembl/drugability/domain/<DOMAIN>>, where <DOMAIN> is the domain number for the hit.

2.8 Discussion

The methods described in this study were motivated by the drug target selection criteria. By bringing together and mining all data that is relevant to the six main target assessment criteria (essentiality, assay feasibility, resistance, toxicity, structural information and druggability) discussed in Section 1.3, and using this data in a species comparative manner, it is possible to identify proteins that are more important to the survival of the parasite which may be used as drug targets for malaria. With the annotation data from the source databases of the protein sequences, data from InterProScan, GO annotation data as well as pathway information, the role played by proteins in the survival and spread of the malaria disease can be assessed. The EC numbers as well as enzyme data can be used to assess proteins on whether they can be readily expressed in laboratories and whether any activity assays have been developed. Such data is useful in obtaining the proteins in their pure forms and in large quantities for HTS, as well as

designing activity detection methods when small molecules are screened against proteins.

Proteins involved in metabolic pathways that do not produce or consume a unique metabolite cannot be good drug targets as there is a potential for resistance. If the target is not unique, or involved in unique reaction, inhibition of the target by drugs will have no effect due to flexibility of pathways. Isoforms or other pathways in the organism may replace the function of the inhibited protein, thus causing resistance to the drug. However, mutations are also an important and major cause of drug resistance. It is therefore important to assess the possibility of drug resistance for a particular protein by analyzing metabolic pathways for choke-points, identifying isoforms, as well as analyzing literature studies on mutations.

Toxicity in humans caused by undesired binding may be assessed with the orthology data from OrthoMCL. Toxicity may be due to a presence of a protein in human that is homologous to a drug target in the parasite. By analyzing the ortholog clusters derived from OrthoMCL, such proteins may be identified. Structural information of the active sites may also be used to assess toxicity by comparing the active sites of the homologous protein in humans to that of the drug target. Having experimentally determined or predicted crystal structures is advantageous to the *in silico* docking techniques as well as during optimization of a drug. These data can be obtained from PDB and MODBASE. It is also important to identify whether a protein has a potential for binding small drug-like molecules with high affinity, a property known as druggability. DrugEBility is one database that provides druggability data. With druggability data, proteins that are more likely to bind small drug-like molecules can be assessed.

Chapter 3

Results and discussion

3.1 Introduction

Discovery was designed for the purpose of mining protein and small molecule data that is relevant to the selection of potential drug targets and small molecule candidates for design of drugs against the malaria disease. The new version of Discovery, Discovery 2.0 (<http://discovery.bi.up.ac.za/>), was developed in Java to speed up the access and retrieval of data as well as to allow for the incorporation of applets that are useful for the analysis of data. It also has added protein annotations which include UniProt accessions and druggability information. The protein annotation data currently available in Discovery 2.0 covers the six *Plasmodium* parasites (*P. berghei*, *P. chabaudi*, *P. falciparum*, *P. knowlesi*, *P. vivax* and *P. yoelii*), the mosquito vector (*A. gambiae*) as well as the human host (*H. sapiens*).

In this chapter, the results from the collection and mining of different data types from a variety of databases and resources for the annotation of proteins are presented together with analyses of proteins done in Discovery 2.0. Firstly, the Discovery 2.0 web interface, along with the different search strategies, will be presented using a potential malaria drug target from *P. falciparum* as an example to demonstrate the different categories of data available and how the data is laid out on the system. This will be followed by the analysis of the annotation data in Discovery 2.0 and case studies on five different protein from *P. falciparum* demonstrating the use of the **advanced search** feature of Discovery 2.0. Lastly, a sample investigation using a known malaria drug target will be presented to demonstrate how the data in Discovery 2.0 can be used to assess a protein as a drug target.

3.2 The Discovery 2.0 web-interface

Users can access the Discovery 2.0 database at <http://discovery.bi.up.ac.za/>. There are four ways in which users can obtain data on a particular protein sequence, i.e. through a **basic search**, **advanced search**, **chemical search** as well as **clinical trials** (Figure 3.1). The last two methods of accessing data will however not be discussed in this study. In the **basic search**, a user may enter a PlasmoDB (in case of parasite protein sequence search) or Ensemble (in case of human or mosquito protein sequence search) protein identifier, a UniProt accession or search by a protein name. A built in auto-complete feature in the protein identifier and UniProt accession searches assists users as they enter characters in the search space (Figure 3.1).

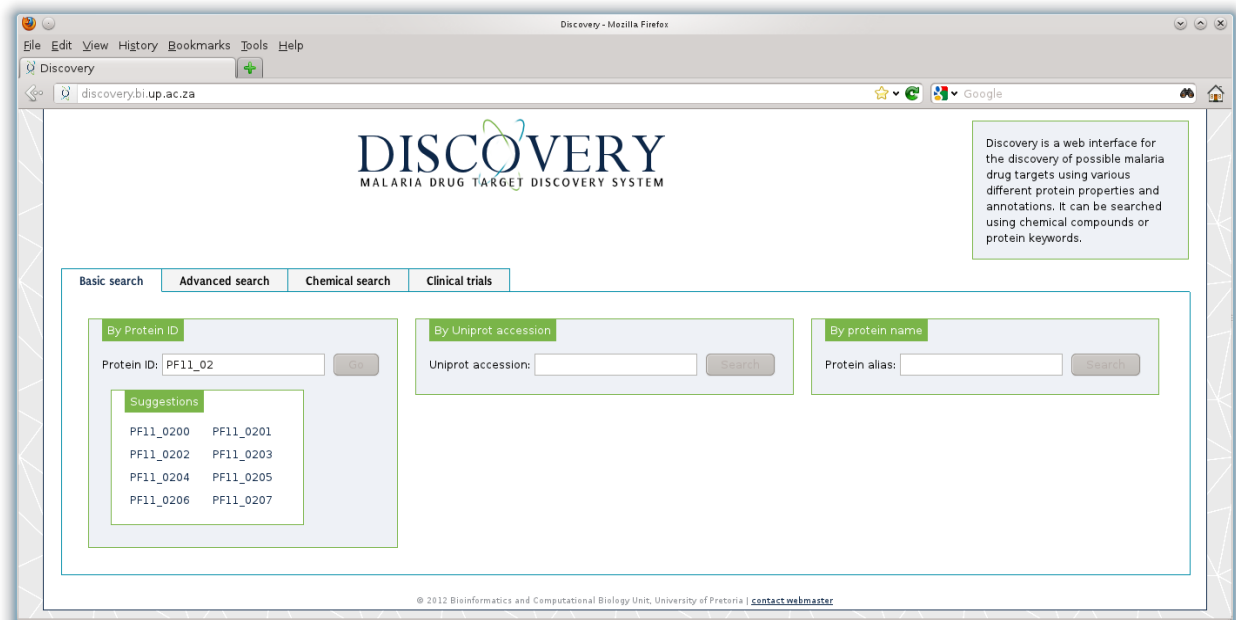


Figure 3.1: **Discovery 2.0 home page.** Entry point to the different types of searches.

With the **advanced search**, users can filter proteins using different search criteria in a step-by-step manner to select the proteins that they are interested in (Figure 3.2). The available search criteria for filtering protein sequences include *function*, *gene ontology*, *MODBASE structures*, *organism*, *orthology*, *PDB-BLAST*, *protein-ligand interactions*, *protein name* and *related PubMed articles*. A list of proteins matching the search criteria is returned, and the user may select the desired protein to view annotations.

To demonstrate the different categories of annotation data available for proteins in Discovery 2.0, a sample investigation was done using a previously described potential target for malaria

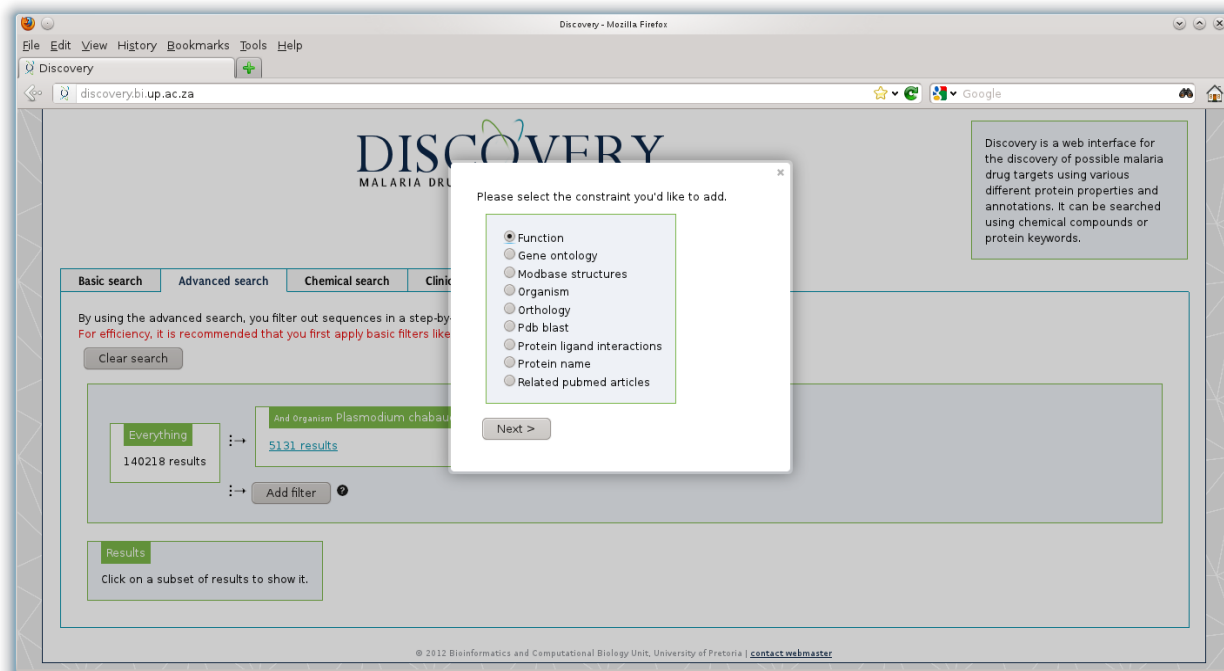


Figure 3.2: **Advanced search.** Allows filtering of sequences based on annotation properties.

(Nguyen *et al.*, 2005). The enzyme deoxyuridine 5'-triphosphate nucleotidohydrolase (dUTPase) from *P. falciparum* (referred to as *Pf*dUTPase from here on), involved in the metabolism of nucleotides, was selected. *Pf*dUTPase catalyzes the hydrolysis of deoxyuridine triphosphate (dUTP) to deoxyuridine monophosphate (dUMP) and pyrophosphate. It does this in the presence of magnesium ions (Figure 3.3). dUTPase is found in eukaryotic and prokaryotic organisms as well as in some viruses. The enzyme has been shown to be an attractive drug target for most organisms since it is crucial for DNA integrity by preventing buildup of dUTP and ensuring the provision of dUMP for the synthesis of deoxythymidine triphosphate (dTTP). This results in low dUTP/dTTP ratios which reduce the misincorporation of uracil into DNA, as DNA polymerases cannot distinguish between dUTP and dTTP (Nguyen *et al.*, 2005, 2006;

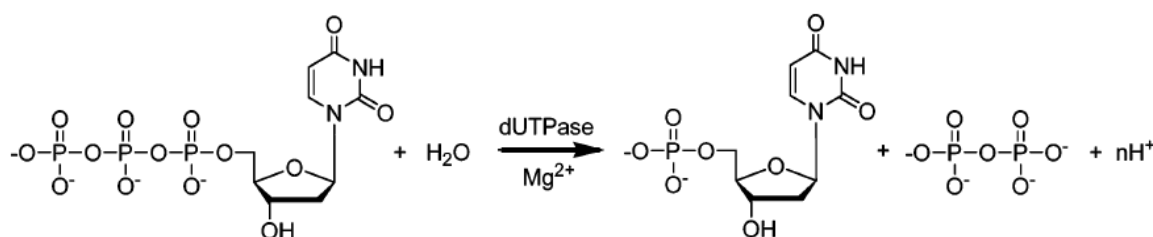


Figure 3.3: **Reaction catalysed by dUTPase.** [Adapted from Nguyen *et al.* (2005)].

Quesada-Soriano *et al.*, 2008). However, the misincorporated uracil in DNA is repaired by DNA glycosylase through excision and replacement with thymine.

When dUTPase is inhibited, the increased levels of dUTP that results cause an increase in the misincorporation of uracil into DNA. The DNA repair process causes high repetitive cycles of excision and replacement of uracil, which lead to increased mutation levels, fragmentation of DNA and eventually cell death (Nguyen *et al.*, 2005; Quesada-Soriano *et al.*, 2008; Vértessy and Tóth, 2009). dUTPases are classified into three families base on their subunit composition, i.e. monomers, dimers and trimers. Whittingham *et al.* (2005) reported a crystal structure of the trimeric *Pfd*UTPase enzyme, each subunit with a substrate binding site complexed with its inhibitor. Small molecule inhibitors have been designed against *Pfd*UTPase, which are mainly analogues of dUMP. These small molecule inhibitors were shown to be selective when tested against the *H. sapiens* dUTPase (*Hsd*UTPase) and also inhibited parasite growth when tested *in vitro* (Nguyen *et al.*, 2005; Whittingham *et al.*, 2005; Nguyen *et al.*, 2006). This showed that *Pfd*UTPase is indeed essential for parasite survival and can be used as a potential drug target for malaria.

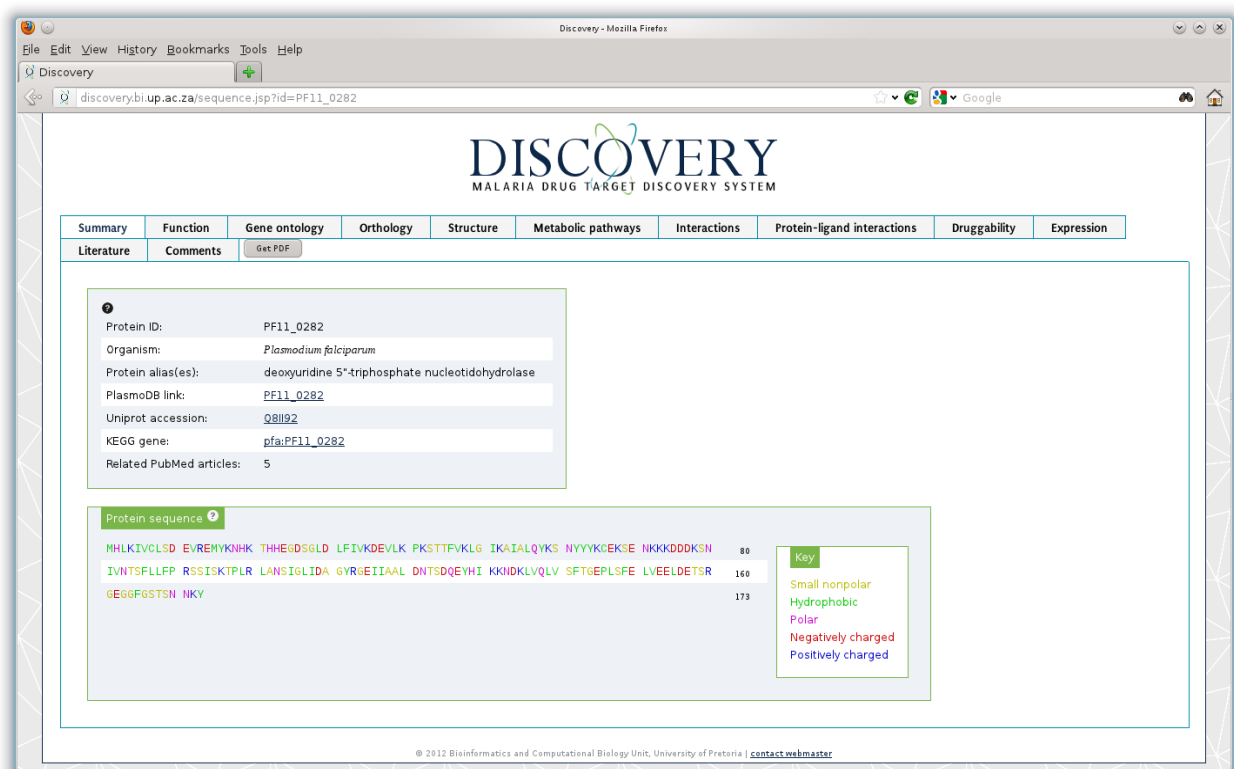
To search for *Pfd*UTPase in Discovery 2.0 using the **basic search**, the user may enter either the enzyme's PlasmDB identifier (PF11_0282), its UniProt accession (Q8II92) or the name of the enzyme. If, however, the user has no information on the protein, but has a set of properties or criteria that should be met in the search, the **advanced search** may be used. For example, since we know *Pfd*UTPase belongs to *P. falciparum*, and that based on literature it has an experimentally determined crystal structure and involved in DNA replication, one could use this information as filters to obtain its annotation data in Discovery using the advanced search. The user would first select an **organism filter** and select *P. falciparum* from the organism list. Once the **organism filter** has been applied, the user may go on to select the **related PubMed articles filter** to get all proteins with at least one PubMed article. After the literature filter has been applied, the **PDB-BLAST filter** may be selected and the desired *E*-value entered. After applying the **PDB-BLAST filter**, the **gene ontology filter** may then be applied, where the user may enter a search term for the desired GO term associated with the protein, like "DNA replication".

From the list of proteins that match the criteria, the user may browse the list and select the

*Pfd*UTPase enzyme to explore the annotation data available. The annotation data for each protein is organized into tabs representing different categories of annotation data (Figure 3.4). These categories are “**Summary**”, “**Function**”, “**Gene ontology**”, “**Orthology**”, “**Structure**”, “**Metabolic pathways**”, “**Protein-ligand interactions**”, “**Druggability**”, “**Expression**” and “**Literature**”. An extra tab, “**Comments**”, is available where users may submit their comments. The data available in these categories will be discussed in the following subsections. The “**Protein-ligand interactions**”, “**Expression**” and “**Literature**” tabs will not be discussed in this study.

3.2.1 Summary

The first tab that shows up in the annotation data is the **summary tab**, displaying the summary information for the protein (Figure 3.4). The protein names and synonyms are displayed in this category together with the protein identifier used in the source database, the UniProt accession and KEGG GENE identifier if available. The number of PubMed articles



The screenshot shows a web browser window displaying the DISCOVERY Malaria Drug Target Discovery System. The interface features a navigation menu with tabs for Summary, Function, Gene ontology, Orthology, Structure, Metabolic pathways, Interactions, Protein-ligand interactions, Druggability, and Expression. The Summary tab is active, showing the following information:

- Protein ID: PF11_0282
- Organism: *Plasmodium falciparum*
- Protein alias(es): deoxyuridine 5'-triphosphate nucleotidohydrolase
- PlasmoDB link: PF11_0282
- Uniprot accession: Q8I192
- KEGG gene: pfa:PF11_0282
- Related PubMed articles: 5

Below this information is the protein sequence, color-coded by properties:

```
MHLKIVCLSD EVREMYKNIHK THHEGDSGLD LFIIVKDEVLK PKSTTFVKLG IKAIALQYKS NYYKCEKSE NKKKDDDKSN 80
IVNTSFLFFP RSSISKPLR LANSIGLIDA GYRGEIIAAL DNTSDQYEHI KKINDKLVQLV SFTGEPLEFSE LVEELDETSR 160
GEGFGSTSN NKY 173
```

A key indicates the color coding: Small nonpolar (green), Hydrophobic (yellow), Polar (purple), Negatively charged (red), and Positively charged (blue).

Figure 3.4: **Summary tab**. Displays the basic information about the protein and displays the amino acid sequence of the protein.

associated with the protein is also shown. Clicking on the protein identifiers takes the user to the corresponding databases for cross-referencing. A graphical representation of the protein sequence is also shown, with amino acids color coded to show small non-polar amino acids (yellow), hydrophobic amino acids (green), polar amino acids (pink), negatively charged amino acids (red) and positively charged amino acids (blue), as shown in Figure 3.4.

3.2.2 Function

The **functions tab** (Figure 3.5) displays the predicted protein families, domains, and functional sites results of the InterProScan. A graphical representation of the InterPro features covering the analyzed protein sequence is shown. A table describing in more details the InterPro analysis is also shown. The table describes the type of analysis used (signature database or tool), the signature identifier, the description of the matching InterPro signature, its start and end positions, the *E*-value score of the match and the corresponding InterPro identifier (Figure 3.5). Clicking on the InterPro identifier takes the user to the InterPro database for a more detailed information on the entry. The user can sort the entries in the table alphabetically or numerically by clicking on the headings of each of the columns.

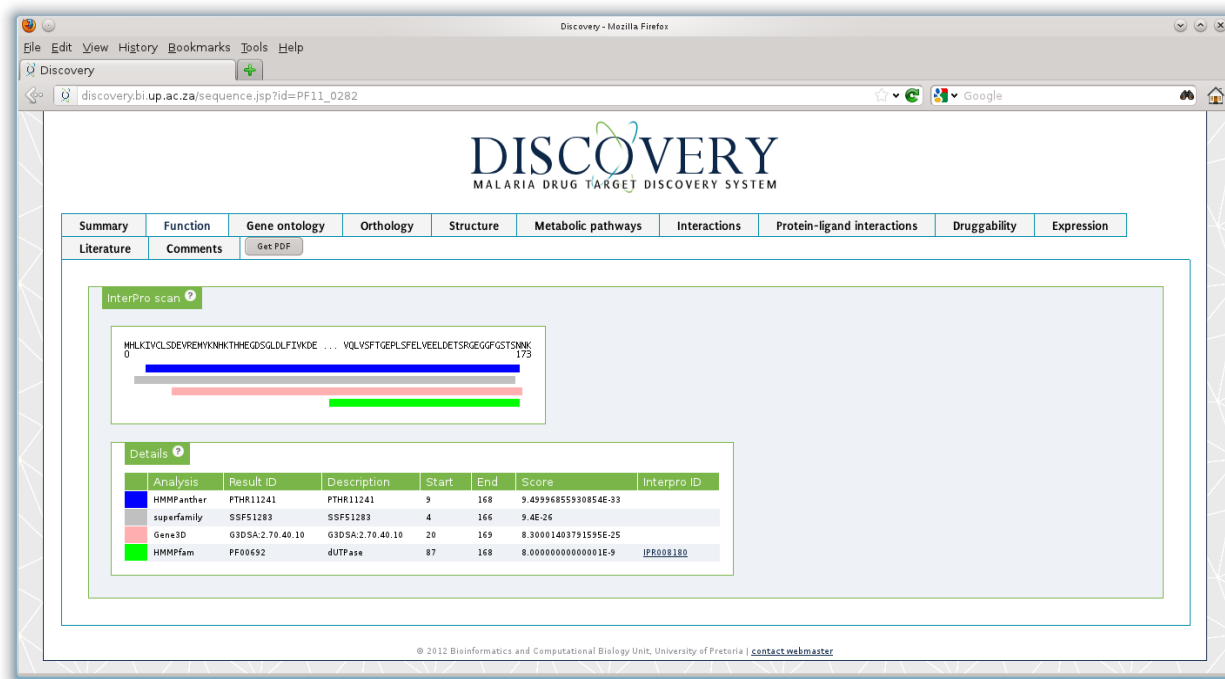


Figure 3.5: **Predicted functions tab.** Displays the results for protein families, domain and functional site analysis for the protein predicted with InterProScan.

3.2.3 Gene Ontology

The **gene ontology tab** displays the GO terms assigned to the protein (Figure 3.6). The GO terms are categorized according to the three domains, i.e. cellular component, molecular function and biological processes. Clicking on each GO term takes the user to the Gene Ontology database for more information on the GO term. Next to each GO term is a “[view]” button, which users may click to view each GO term in graph form using the AmiGO visualization tool.

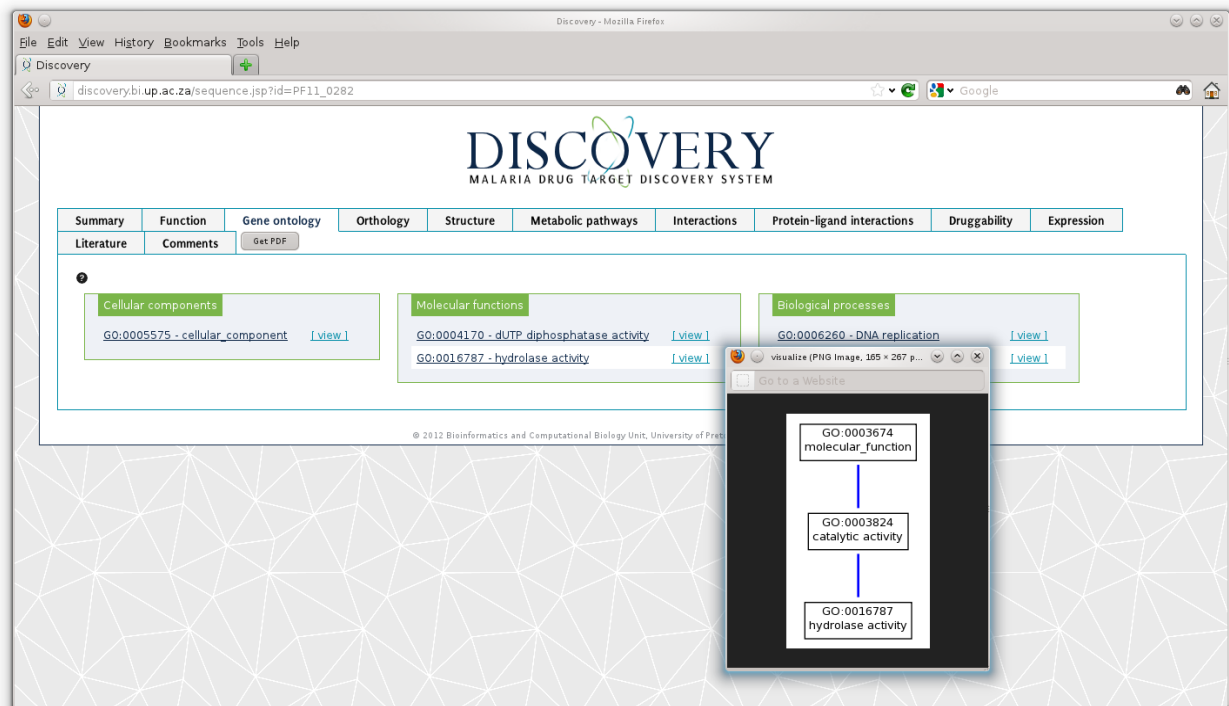
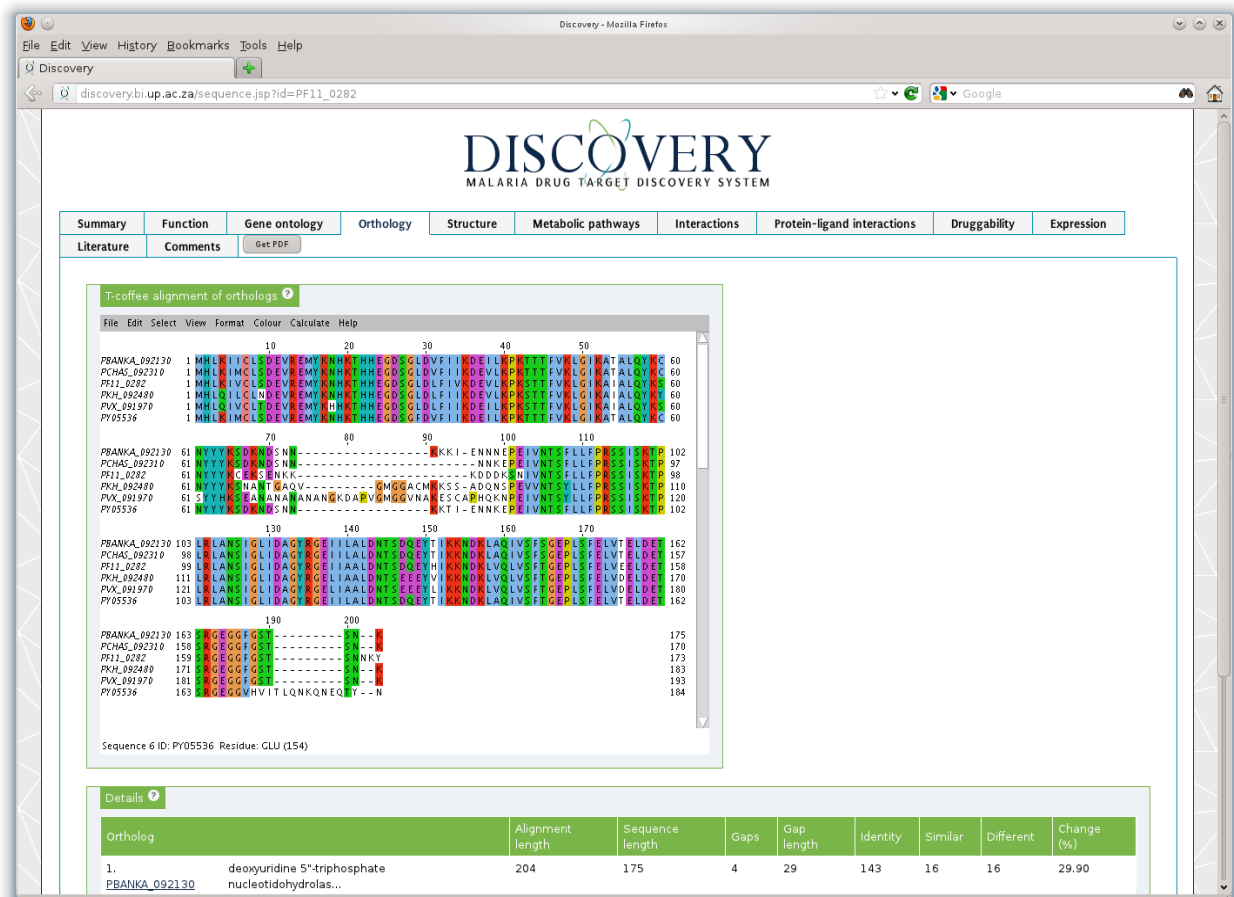


Figure 3.6: **Gene Ontology tab**. Displays the results for GO annotation, with the GO terms assigned to the query protein classified according to the three domains (cellular component, molecular function and biological processes).

3.2.4 Orthology

The results from the OrthoMCL clustering are displayed in the **orthology tab** (Figure 3.7). In this tab, the multiple sequence alignment of the protein sequences that form the cluster is displayed in the Jalview applet, with the default Clustalx color coding for the amino acids. Users can interact with and manipulate the alignment data using the Jalview applet. A table displaying the details of the sequences in the multiple sequence alignment is also presented in the orthology tab. The table was generated with InfoAlign and describes the names of the proteins in the alignment, the alignment and the sequence lengths, the gaps and gap lengths, the number

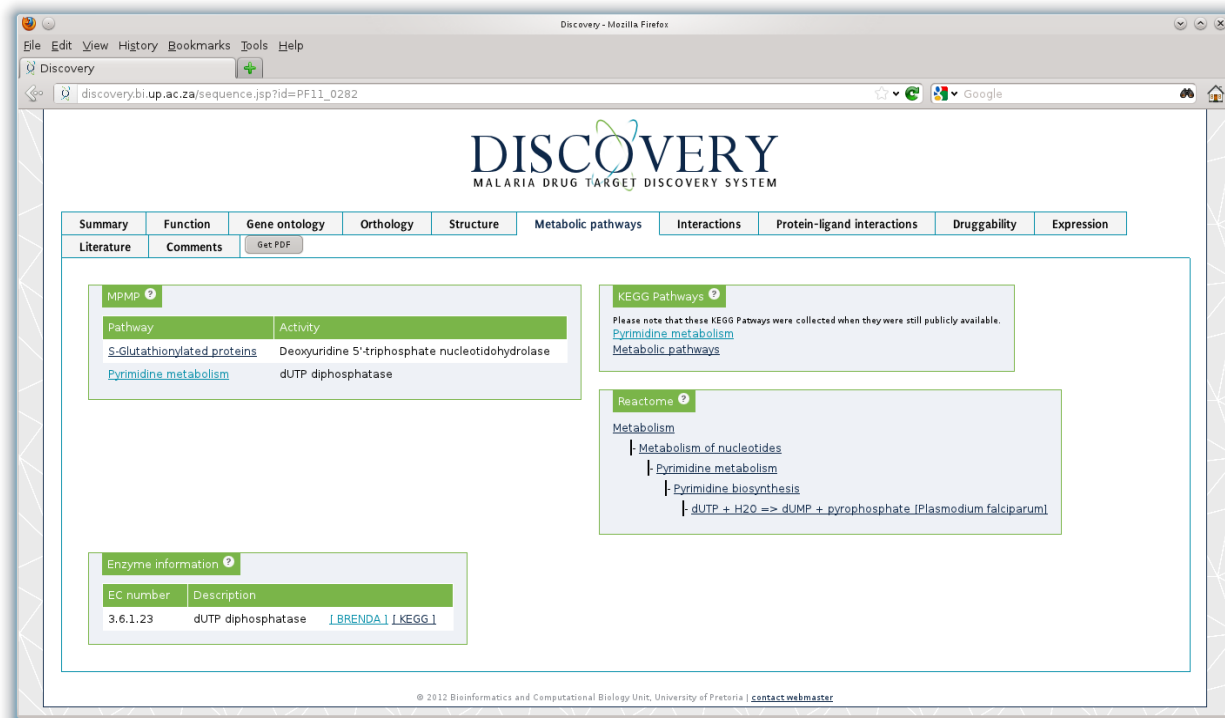


of identical, similar and different residues in each sequence compared to the reference sequence (the query sequence, PF11_0282 in this case) as well as the percentage change between the sequence and the reference sequence (query sequence). The results in the table can also be sorted alphabetically or numerically by clicking on the headings in the columns. Clicking on each of the proteins identifiers in the table opens a Discovery 2.0 page for that protein entry in a new window.

3.2.5 Metabolic pathways

The **metabolic pathways tab** displays the MPMP, KEGG and Reactome pathways that the protein is involved in as well as the enzyme information, if available (Figure 3.8). Pathways are separated according to the source database that they were obtained from. Clicking on the name of the pathway takes the user to the respective database for a full view of the pathways.

The enzyme information table displays the EC number for the protein, a description of the enzyme as well as the links to BRENDA and KEGG ENZYME.



The screenshot shows a web browser window displaying the DISCOVERY Malaria Drug Target Discovery System. The 'Metabolic pathways' tab is selected, showing a table of pathways and enzyme information.

Summary	Function	Gene ontology	Orthology	Structure	Metabolic pathways	Interactions	Protein-ligand interactions	Druggability	Expression						
Literature	Comments	Get PDF													
MPMP <table border="1"> <thead> <tr> <th>Pathway</th> <th>Activity</th> </tr> </thead> <tbody> <tr> <td>S-Glutathionylated proteins</td> <td>Deoxyuridine 5'-triphosphate nucleotidohydrolase</td> </tr> <tr> <td>Pyrimidine metabolism</td> <td>dUTP diphosphatase</td> </tr> </tbody> </table>		Pathway	Activity	S-Glutathionylated proteins	Deoxyuridine 5'-triphosphate nucleotidohydrolase	Pyrimidine metabolism	dUTP diphosphatase	KEGG Pathways Please note that these KEGG Pathways were collected when they were still publicly available. Pyrimidine metabolism Metabolic pathways							
Pathway	Activity														
S-Glutathionylated proteins	Deoxyuridine 5'-triphosphate nucleotidohydrolase														
Pyrimidine metabolism	dUTP diphosphatase														
Enzyme information <table border="1"> <thead> <tr> <th>EC number</th> <th>Description</th> </tr> </thead> <tbody> <tr> <td>3.6.1.23</td> <td>dUTP diphosphatase BRENDA KEGG</td> </tr> </tbody> </table>		EC number	Description	3.6.1.23	dUTP diphosphatase BRENDA KEGG	Reactome Metabolism <ul style="list-style-type: none"> Metabolism of nucleotides Pyrimidine metabolism Pyrimidine biosynthesis dUTP + H2O => dUMP + pyrophosphate [Plasmodium falciparum] 									
EC number	Description														
3.6.1.23	dUTP diphosphatase BRENDA KEGG														

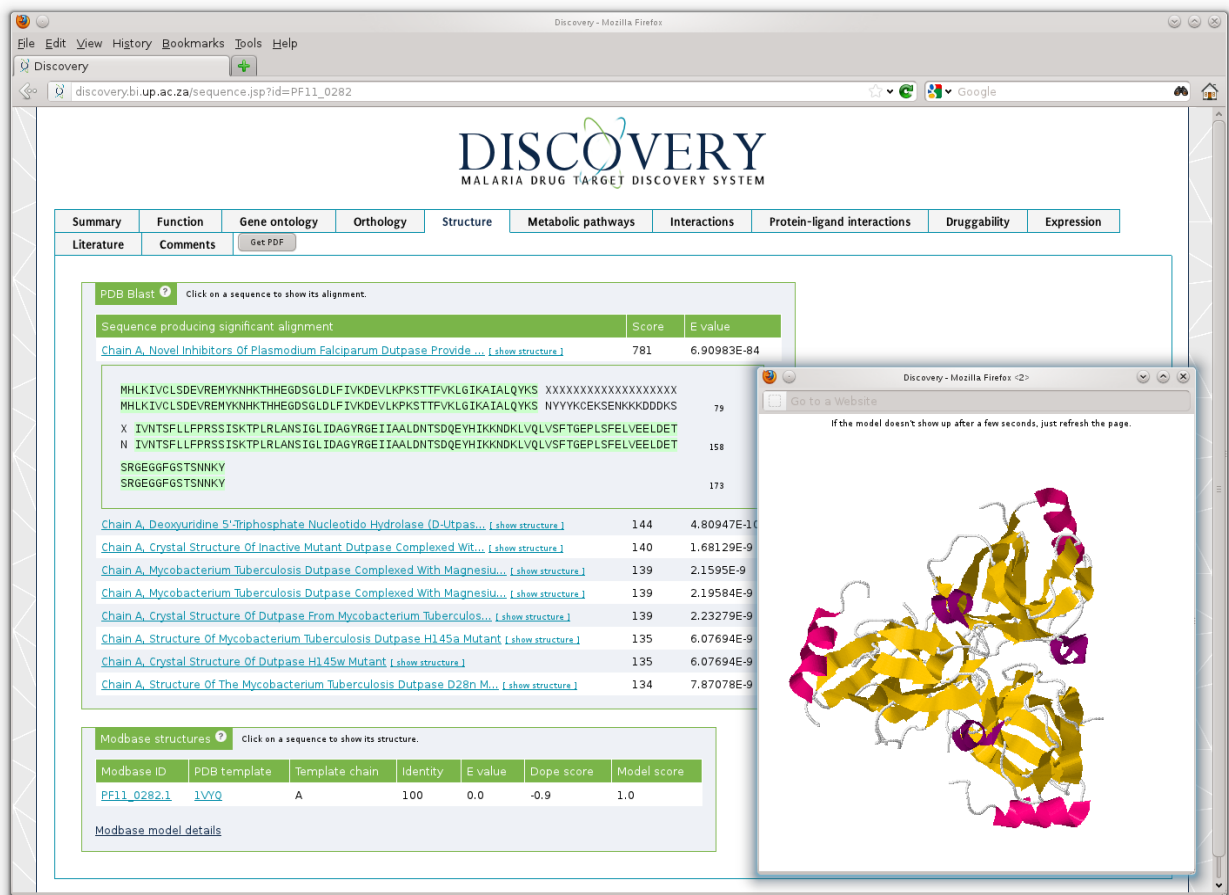
© 2012 Bioinformatics and Computational Biology Unit, University of Pretoria | [contact webmaster](#)

Figure 3.8: **Metabolic pathways tab.** Shows the EC numbers and the different pathways that the query protein is involved in.

3.2.6 Structure

The **structure tab** displays the results of the PDB-BLAST as well as the predicted crystal structures for a protein from the MODBASE database (Figure 3.9). The experimentally determined structures in this tab are evident by the low BLAST *E*-values (usually zero) along with the hits that are similar to the protein. Clicking on the PDB-BLAST hit displays the sequence alignment between the hit and the query protein. Clicking on the “[show structure]” button displays a pop-up window showing the crystal structure in the Jmol application.

The predicted MODBASE structures are displayed in a table showing statistics of the predictions (Figure 3.9). The table shows the MODBASE identifier, the PDB template and chain used to determine the structure, the sequence identity, the *E*-value of the alignment between the query protein and the template, the atomic distance-dependent DOPE score and the model score (reliability of the model). Clicking on the MODBASE identifier brings up a pop-up window with the predicted crystal structure in Jmol. Clicking on the PDB template also displays



DISCOVERY
MALARIA DRUG TARGET DISCOVERY SYSTEM

Summary Function Gene ontology Orthology Structure **Crystal structures** Metabolic pathways Interactions Protein-ligand interactions Druggability Expression

Literature Comments [Get PDF](#)

PDB Blast [Click on a sequence to show its alignment.](#)

Sequence producing significant alignment	Score	E value
Chain A, Novel Inhibitors Of Plasmodium Falciparum Dufpase Provide ... [show structure]	781	6.90983E-84
MHLKIVCLSDVEVREHYKHKHTHEGDSGLDLFIVKDEVLKPKSTTFVKLGIIKAIALQYKS XXXXXXXXXXXXXXXXXXXX		79
MHLKIVCLSDVEVREHYKHKHTHEGDSGLDLFIVKDEVLKPKSTTFVKLGIIKAIALQYKS NYYYYKEKSENKKDDDKS		
X IVNTSFLFPRSSISKTPRLRLANSIGLIDAGYRGEIIAALDNTSDQEYHIKNDKLVQVLSFTGPELSFELVEELDET		158
N IVNTSFLFPRSSISKTPRLRLANSIGLIDAGYRGEIIAALDNTSDQEYHIKNDKLVQVLSFTGPELSFELVEELDET		
SRGEGFGSTSNINKY		173
SRGEGFGSTSNINKY		
Chain A, Deoxyuridine 5'-Triphosphate Nucleotido Hydrolase (D-Utpas... [show structure]	144	4.80947E-11
Chain A, Crystal Structure Of Inactive Mutant Dufpase Complexed Wit... [show structure]	140	1.68129E-9
Chain A, Mycobacterium Tuberculosis Dufpase Complexed With Magnesi... [show structure]	139	2.1595E-9
Chain A, Mycobacterium Tuberculosis Dufpase Complexed With Magnesi... [show structure]	139	2.19584E-9
Chain A, Crystal Structure Of Dufpase From Mycobacterium Tuberculos... [show structure]	139	2.23279E-9
Chain A, Structure Of Mycobacterium Tuberculosis Dufpase H145a Mutant [show structure]	135	6.07694E-9
Chain A, Crystal Structure Of Dufpase H145w Mutant [show structure]	135	6.07694E-9
Chain A, Structure Of The Mycobacterium Tuberculosis Dufpase D28n M... [show structure]	134	7.87078E-9

Modbase structures [Click on a sequence to show its structure.](#)

Modbase ID	PDB template	Template chain	Identity	E value	Dope score	Model score
PF11_0282.1	1VYQ	A	100	0.0	-0.9	1.0

[Modbase model details](#)

Figure 3.9: **Crystal structures** tab. Displays the PDB-BLAST results for the query protein along with the predicted structures from MODBASE.

the crystal structure in Jmol. A link to MODBASE is available so that the user may browse the model predictions in more details. The results for both PDB-BLAST and MODBASE structures can be sorted by clicking on the headings of the columns.

3.2.7 Interactions

The **interactions** tab displays the protein-protein interactions between the query protein and other proteins (Figure 3.10). The interactions are grouped as obtained from each of the three interaction databases used in this study (IntAct, MINT and DIP). The interaction data shows the UniProt accessions and names for both the interactors, the interaction detection method used, the type of interaction, the taxonomy for both interactors as well as the identifier for the interaction used in the parent database. Clicking on the column headings sorts the results numerically or alphabetically. Clicking on the UniProt accessions takes the user to the UniProt

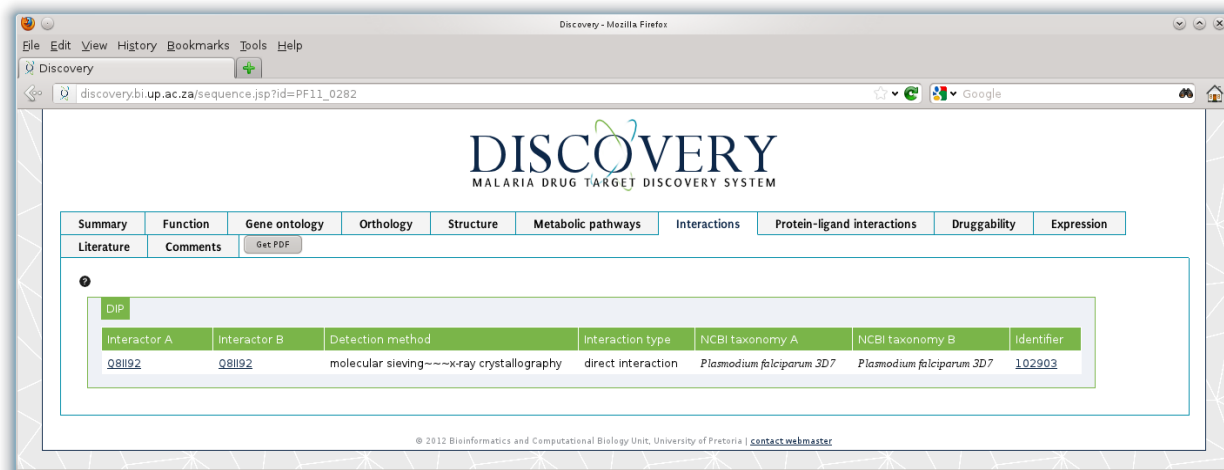


Figure 3.10: **Interactions tab.** Displays the proteins that interact with the query protein, the type of interaction as well as the method used for detection.

database to view the protein entry and clicking on the interaction identifier takes the user to the parent database for more details on the interaction. In cases where an interaction score is present, it is displayed in the interaction data.

3.2.8 Druggability

The **druggability tab** displays the DrugEBility BLAST results of the domains and gene sequences (Figure 3.11). The results are split into two categories displaying the druggability



Figure 3.11: **Druggability tab.** Shows the significant BLAST hits against the DrugEBility database for the query protein.

of domains and genes. Both tables display the hit sequences, the alignment score and the *E*-value of the alignment. The table of domain druggability has an added column displaying whether the domain is druggable (tick) or undruggable (cross). The results may be sorted alphabetically or numerically by clicking on the column headings in both the tables. Clicking on the sequence identifier (PDB code or UniProt accession) of a hit displays the alignment of the match. Clicking on the “[show domain]” or “[show gene]” take the user to the ChEMBL’s DrugEBIity resource for more details on the druggability calculations of the hit.

3.3 The annotation data in Discovery 2.0

Discovery 2.0 currently contains 140 218 protein sequences, of which 14 324 belong to *A. gambiae*, 92 012 belong to *H. sapiens*, 4 904 belong to *P. berghei*, 5 131 belong to *P. chabaudi*, 5 491 belong to *P. falciparum*, 5 197 belong to *P. knowlesi*, 5 435 belong to *P. vivax* and 7 724 belong to *P. yoelii*. Analysis of the annotation data reveals that 66% (9 389) of the *A. gambiae* and 82% (75 214) of the *H. sapiens* protein sequences were assigned with UniProt accessions (Figure 3.12). In the parasite proteomes, 52% (2 534) of the *P. berghei*, 36% (1 863) of the *P. chabaudi*, 91% (5 018) of the *P. falciparum*, 98% (5 095) of the *P. knowlesi*, 99% (5 380) of the *P. vivax* and 92% (7 116) of the *P. yoelii* protein sequences were assigned with UniProt accession numbers.

Using the UniProt accessions, 39% (5 532) of the *A. gambiae* and 56% (51 939) of the *H. sapiens* protein sequences were assigned with at least one GO term. 50% (2 763) of the *P. falciparum*, 43% (2 257) of the *P. knowlesi*, 42% (2 259) of the *P. vivax* and 28% (2 176) of the *P. yoelii* protein sequences had at least GO term. There was no GO data for the *P. berghei* and *P. chabaudi* as seen in Figure 3.12. The data for both the *P. berghei* and *P. chabaudi* seems to lack not only in GO annotations, but also in metabolic pathway assignments. Only 4 and 6 of the *P. berghei* and *P. chabaudi* protein sequences, respectively, are involved in at least one metabolic pathway. However, the pathway assignment of proteins across all the species in Discovery 2.0 seems to be relatively low. Only 11% (1 599) of the *A. gambiae* and 10% (8 803) of the *H. sapiens* protein sequences were assigned a role in the metabolic pathways. *P. falciparum* had 16% (881) of its proteins assigned to pathways, which was the highest of the parasite proteomes. 15% (768) of the *P. knowlesi*, 14% (758) of the *P. vivax* and 9% (666) of

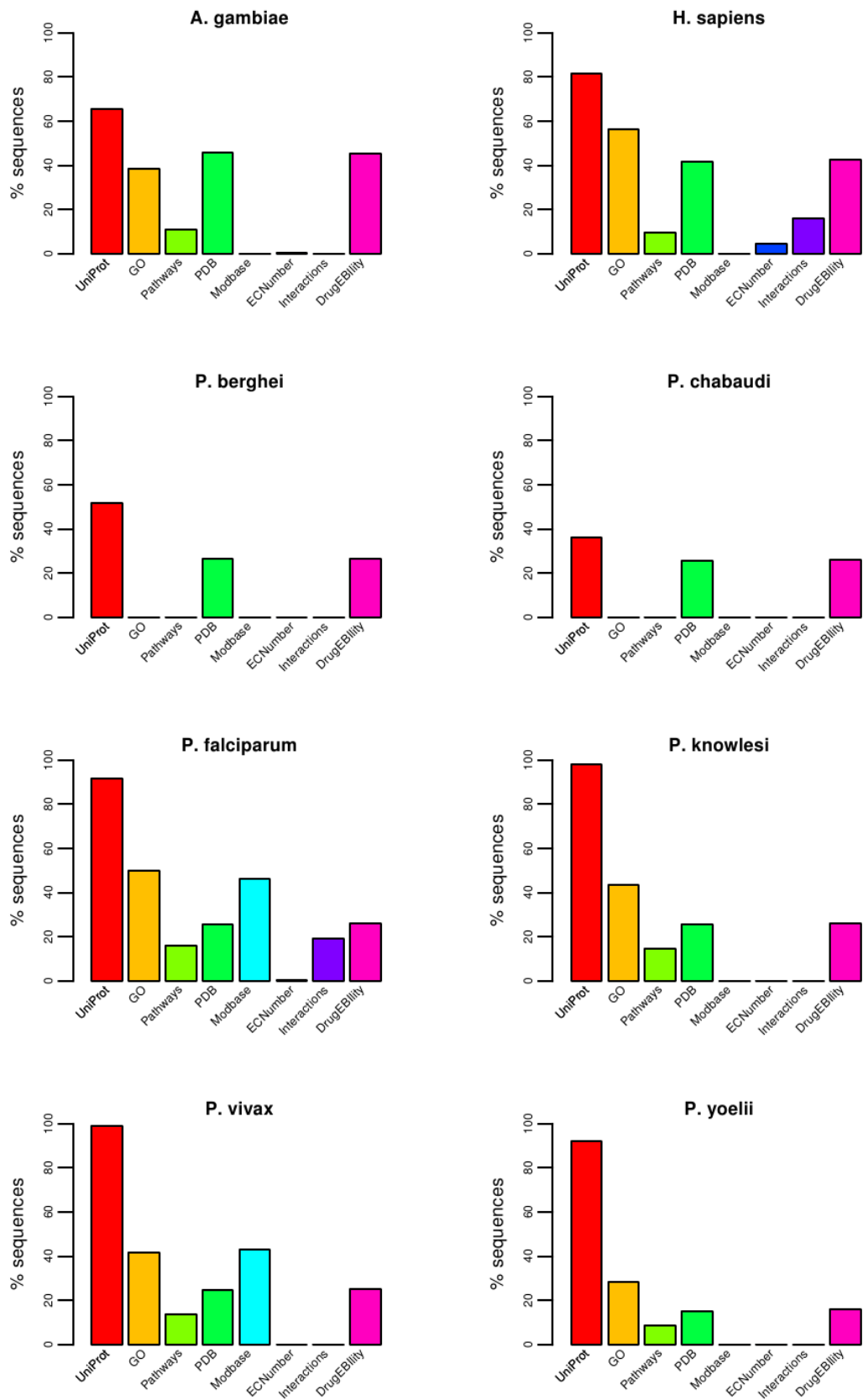


Figure 3.12: **Genome annotations.** The bar graphs show the summary of the annotation of proteins from each of the eight different species in Discovery 2.0. The type of annotation data is shown on the x-axis and the y-axis is the percentage of the sequences annotated for a species (Image generated in using R).

the *P. yoelii* proteins had pathways assigned to them (Figure 3.12).

Forty-six percent (6 585) of the *A. gambiae* protein sequences had at least one significant match with the protein sequences in the PDB database, whilst *H. sapiens* had 42% (38 412). All the parasite species had approximately 25% of their protein sequences producing significant matches with PDB proteins sequences, with the exception of *P. yoelii* which had 15% (1169) of it proteins having at least one matching protein in PDB. The predicted MODBASE models were only available for *H. sapiens*, *P. falciparum* and *P. vivax*, of which 0.01% (11), 46% (2 544), and 43% (2 347) of their proteins were assigned MODBASE models respectively.

The protein sequences in Discovery 2.0 were poorly annotated with respect to EC numbers and protein-protein interactions. Only 45 of the *A. gambiae* and 4 099 (4%) of the *H. sapiens* proteins had EC numbers assigned. Both *P. berghei* and *P. chabaudi* had only 2 of their proteins assigned EC numbers. *P. falciparum* had 21 of its protein sequences assigned EC numbers, whilst *P. knowlesi*, *P. vivax* and *P. yoelii* had 5, 4 and 11 of their protein sequences assigned EC numbers respectively. No protein-protein interaction data was observed for all species except for *H. sapiens* and *P. falciparum* which had 16% (14 898) and 19% (1 053) of their protein having at least one interactor assigned. The number of protein sequences in each of the species having a significant match with at least one protein from the DrugEBIity database is approximately the same as that of the PDB matches (Figure 3.12).

Assigning the proteins in Discovery 2.0 with UniProt accessions was an important step in the annotation of proteins since most databases use UniProt accessions for the data they produce. Having UniProt accessions assigned to proteins made it possible to map different datasets from these databases onto the protein sequences in Discovery 2.0. Although most of the proteins were successfully assigned with UniProt accessions, the *P. berghei* and *P. chabaudi* protein sequences were the most poorly assigned. This shows the lack of studies done on both these species, especially since no GO data was found in the UniProt-GOA database. However, analyzing the GO annotation data in the PlasmoDB database for *P. berghei* and *P. chabaudi* shows that GO terms have been assigned to both the species. *P. berghei* has 2 167 protein sequences with assigned GO terms whilst *P. chabaudi* has 2 182. The reason for this is that GO associations for these species are obtained elsewhere in the PlasmoDB database; thus the links provided in Discovery 2.0 to PlasmoDB ensure that the user has an alternative access to

the data that is missing in Discovery 2.0, if it exists.

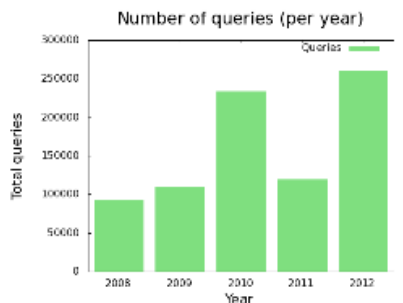
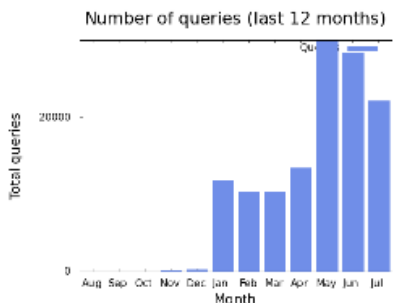
UniProt accessions were also used in metabolic pathway assignment from Reactome and KEGG identifiers were used to assign pathways from KEGG PATHWAY to proteins. For *Plasmodium* species, PlasmoDB was also used to extract pathway information for proteins. However, a very low pathway assignment to proteins of less than 20% is seen throughout the species. For *P. berghei* and *P. chabaudi*, the lack of UniProt accessions and the use of different protein identifiers (protein identifiers are different from those in the KEGG database) explains the reason for the very low pathway assignment. Another reason for the low metabolic pathway assignment is that the pathways from Reactome and MPMP were not included in the analysis since they are fetched from the Reactome and PlasmoDB databases, respectively, as the user accesses the page for a particular protein in Discovery 2.0. The pathways assigned to proteins in the analysis (Figure 3.12) correlate with metabolic pathway data downloaded from KEGG PATHWAYS, which is stored in the system, and thus were the only pathways included in the analysis.

The DrugEBility resource calculates druggability for proteins by using crystal structures from PDB as discussed in Section 1.3. The reason for the similar BLAST searches against PDB and DrugEBility is that both BLAST searches are basically done on the same protein sequences of the crystal structures from PDB. The MODBASE models were successfully mapped to the proteins from *P. falciparum* and *P. vivax*, as the data correlates with the statistics from the MODBASE database (Figure 3.13). However, the statistics also reveal that the mapping of *H. sapiens* protein sequences to their MODBASE models was unsuccessful. Even though there were fewer than expected models mapped to the *H. sapiens* proteins, the links to the MODBASE database available in Discovery 2.0 provide an alternative for obtaining the predicted structure information for proteins. Not many *Plasmodium* proteins were assigned EC numbers using the enzyme data file downloaded from ExPASy. As an alternative, the EC numbers for *Plasmodium* species proteins were obtained from PlasmoDB. However, even with the EC numbers being extracted from PlasmoDB, the number of proteins assigned EC numbers was still very low as seen in Figure 3.12.



[Sali Lab Home](#) [ModWeb](#) [ModLoop](#) [ModBase](#) [ModEval](#) [PCSS](#) [FoXS](#) [IMP](#) [ModPipe](#)

ModBase Access



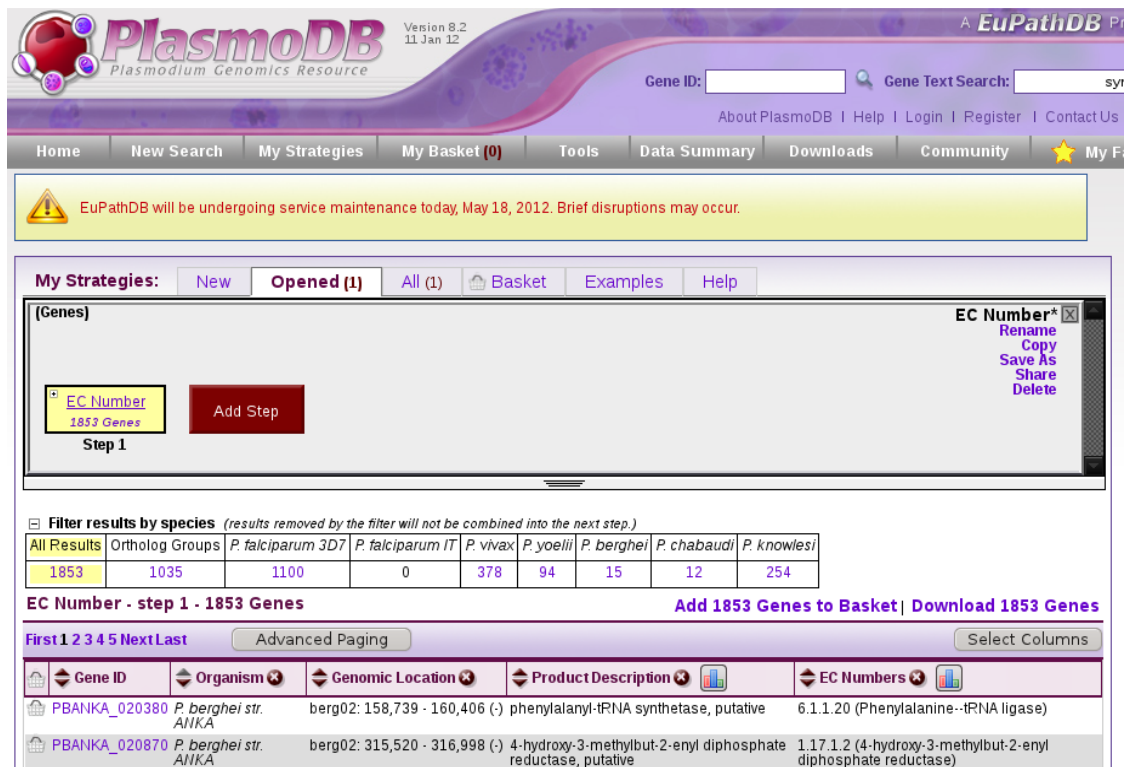
ModBase Contents

Number of Models 24,672,117
 Number of Unique Sequences modeled 3,811,465
 Unique sequences attempted 5,037,714
 Number of PDB chains 47,115
 Statistics updated at Fri Jul 20 22:03:52 PDT 2012

Genome Datasets:

Genome	Number of Transcripts	Number of modeled Transcripts	Number of Models
Archaeoglobus fulgidus	2,409	1,794	3,980
Bacillus subtilis	4,105	3,373	9,244
Brugia Malayi	11,397	7,850	23,216
Burkholderia mallei	4,798	3,908	11,032
Caenorhabditis elegans	22,698	18,996	52,232
Canis familiaris	30,264	22,614	65,595
Clostridium tetani	2,412	2,158	5,864
Coxsackievirus B3 (strain Woodruff)	15	15	13
Cryptosporidium hominis	3,886	1,614	3,287
Cryptosporidium parvum	3,806	1,918	3,969
Denque virus 2 16681-PDK53	12	12	12
Drosophila melanogaster	17,104	9,381	24,683
Escherichia coli	4,206	3,150	5,994
Helicobacter pylori 26695	1,552	1,528	4,744
Homo sapiens	32,010	21,270	51,076
Human enterovirus 71	25	23	49
Human poliovirus 1 Mahoney	11	11	10
Leishmania major	8,009	3,975	8,285
Methanobrevibacter ruminantium M1	2,209	1,745	3,986
Methanococcus jannaschii	1,785	1,480	1,707
Methanopyrus kandleri	1,687	1,111	2,466
Mus musculus	30,133	25,337	70,765
Mycobacterium leprae	1,601	1,178	2,493
Mycobacterium tuberculosis	3,954	2,808	5,913
Mycoplasma pneumoniae	687	426	857
Nanoarchaeum equitans	536	447	496
Picrophilus torridus	1,535	1,260	2,902
Plasmodium falciparum	5,342	2,599	5,053
Plasmodium vivax	5,334	2,359	4,670
Pseudomonas aeruginosa	5,559	3,806	9,222

Figure 3.13: MODBASE statistics for genomes with modelled structures. 21 270 of the *H. sapiens* proteins have modelled structures whilst *P. falciparum* and *P. vivax* have 2 599 and 2 359 proteins with predicted structures respectively (<http://modbase.compbio.ucsf.edu/modbase-cgi/display.cgi?server=modbase&type=statistics>).



PlasmoDB Version 8.2 11 Jan 12
 Plasmodium Genomics Resource

Gene ID: Gene Text Search:

Home New Search My Strategies My Basket (0) Tools Data Summary Downloads Community My F

EuPathDB will be undergoing service maintenance today, May 18, 2012. Brief disruptions may occur.

My Strategies: New Opened (1) All (1) Basket Examples Help

(Genes) EC Number*
 Rename Copy Save As Share Delete

EC Number 1853 Genes Add Step Step 1

Filter results by species (results removed by the filter will not be combined into the next step.)

All Results	Ortholog Groups	<i>P. falciparum</i> 3D7	<i>P. falciparum</i> IT	<i>P. vivax</i>	<i>P. yoelii</i>	<i>P. berghei</i>	<i>P. chabaudi</i>	<i>P. knowlesi</i>
1853	1035	1100	0	378	94	15	12	254

EC Number - step 1 - 1853 Genes Add 1853 Genes to Basket | Download 1853 Genes

First 1 2 3 4 5 Next Last Advanced Paging Select Columns

Gene ID	Organism	Genomic Location	Product Description	EC Numbers
PBANKA_020380	<i>P. berghei</i> str. ANKA	berg02: 158,739 - 160,406 (-)	phenylalanyl-tRNA synthetase, putative	6.1.1.20 (Phenylalanine-tRNA ligase)
PBANKA_020870	<i>P. berghei</i> str. ANKA	berg02: 315,520 - 316,998 (-)	4-hydroxy-3-methylbut-2-enyl diphosphate reductase, putative	1.17.1.2 (4-hydroxy-3-methylbut-2-enyl diphosphate reductase)

Figure 3.14: Search for proteins by EC numbers in PlasmoDB. Results from a search in PlasmoDB to identify the number of proteins assigned with EC numbers in each of the *Plasmodium* species.

A search for proteins assigned with EC numbers was conducted in PlasmoDB by selecting all organisms and using a “*” as a wild-card to find all protein with EC numbers. The results (Figure 3.14) showed that the number of proteins assigned with EC numbers is far higher than the one seen in our analysis of annotation data in Discovery 2.0. The reason for this difference is that the *P. falciparum* proteins assigned with EC numbers were missed in the analysis since the EC numbers for *Plasmodium* proteins are fetched from the PlasmoDB database as a user accesses the annotation data for *P. falciparum* proteins in Discovery 2.0, and therefore the data is not stored in the system as with the pathway data from Reactome and MPMP.

3.4 Case studies on Discovery 2.0

The large amounts of protein sequences from the eight different species in Discovery 2.0 it makes difficult for users to find proteins they are interested in if there is no protein identifier or UniProt accession. However, as mentioned before (Section 3.2), the **advanced search** functionality of Discovery 2.0 makes it easier for users to find proteins they are interested in using a set of criteria to be met. The different types of filters available in the **advanced search**

functionality of Discovery 2.0 are shown in Table 3.1 along with their descriptions.

With the filters mentioned in Table 3.1, users may choose to combine multiple filters using the “AND” option or exclude results using the “AND NOT” option on the parameters of each filter. In this section, the advanced search functionality of Discovery 2.0 will be used to demonstrate how users can obtain protein sequences they are interested in using a set of criteria to be met. Five different types of proteins from different protein families were identified through the **advanced search** and their annotation data analyzed. These proteins were: protein kinase, G protein-coupled receptor (GPCR), peptidase, aminopeptidase and a dehydrogenase.

Table 3.1: **Different types of filters available on the advanced search of Discovery 2.0.**

Filter	Description
Function	Filters protein sequences based on their predicted families, domains and sites from the InterProScan results. A search term is entered and matched to the descriptions/domain signatures in the InterProScan results.
Gene ontology	Filters protein sequences based on the GO term descriptions. A search term is entered and matched against the GO terms in the Discovery 2.0 database.
MODBASE structures	Filters protein sequences based on the presence or absence of at least one predicted MODBASE structure in the database.
Organism	Filters protein sequences based on the organism(s) selected.
Orthology	Filters protein sequences based on the presence of orthologous proteins in other species (orthologs) or within a species (paralogs). A desired organism may be selected and a %change parameter (measure of how different an ortholog is to a reference sequence) set to filter out the sequences.
PDB-BLAST	Filters protein sequences based on the BLAST search results against the PDB database. A desired <i>E</i> -value cut-off may be set with this filter to match against the PDB-BLAST results in the database.
Protein-ligand interactions ^a	Filters protein sequences based on the BLAST results against the ChEMBL database. A desired <i>E</i> -value cut-off may be set and a search term to match with small molecule names in the database entered.
Protein name	Filters protein sequences based on the names of proteins in the database. A search term is entered and matched against the protein names in the database.
Related PubMed articles	Filters protein sequences based on the presence or absence of at least one PubMed article associated with proteins in the database.

^aThe details of this filter fall beyond the scope of this thesis and therefore will not be discussed here.

3.4.1 Protein kinase

Most eukaryotic cellular processes are regulated by protein kinases and phosphatases through reversible protein phosphorylation i.e., protein kinases add phosphate groups to substrate proteins, whilst phosphatases are responsible for the removal of the phosphate group on proteins (Manning *et al.*, 2002; Ward *et al.*, 2004). In eukaryotes, protein kinases are amongst the largest family of genes, which can be classified into two groups: the “conventional” eukaryotic protein kinases (ePK) and “atypical” protein kinases (aPK) (Manning *et al.*, 2002; Ward *et al.*, 2004). ePKs are the largest group and are further sub-classified into eight families based on their domains and modes of regulation. aPK’s on the other hand do not share any similarity with ePKs but have been shown to have protein kinase activity (Miranda-Saavedra and Barton, 2007). Protein kinases are attractive drug targets because of their role in essential cellular processes; and also because mutations and dysregulation of protein kinases results in many human diseases (Hopkins and Groom, 2002; Ward *et al.*, 2004).

To identify a protein sequence belonging to the protein kinase superfamily in Discovery 2.0 using the **advanced search**, a set of properties were set to guide the search. The protein kinase sequence of interest must be from *P. falciparum* and must have some related PubMed articles to show that some research has been done on the protein. The protein must also have significant BLAST hits against PDB or have at least one predicted MODBASE structure present. By adding the *organism filter* first on the **advanced search tab** in Discovery 2.0, the results were reduced from 140 218 sequences to 5 491 sequences which only belong to *P. falciparum* (Figure 3.15). Applying the *related PubMed articles filter*, the *MODBASE structures filter* as well as the *PDB-BLAST filter* (default *E*-value), the results were further reduced to 231 protein sequences.

However, 231 protein sequences was still a lot to work with and protein sequences not belonging to the protein kinase family were also included. The more advanced *gene ontology* and *function filters* were used to exclude all proteins that were not kinases. Protein kinases are involved in the process of phosphorylation of other proteins and according to Manning *et al.* (2002), most kinases contain the protein kinase catalytic domain. To identify whether “protein phosphorylation” is a valid GO term, a GO term search was conducted in the Gene Ontology database (<http://www.geneontology.org/>). The results revealed that “protein phosphoryla-

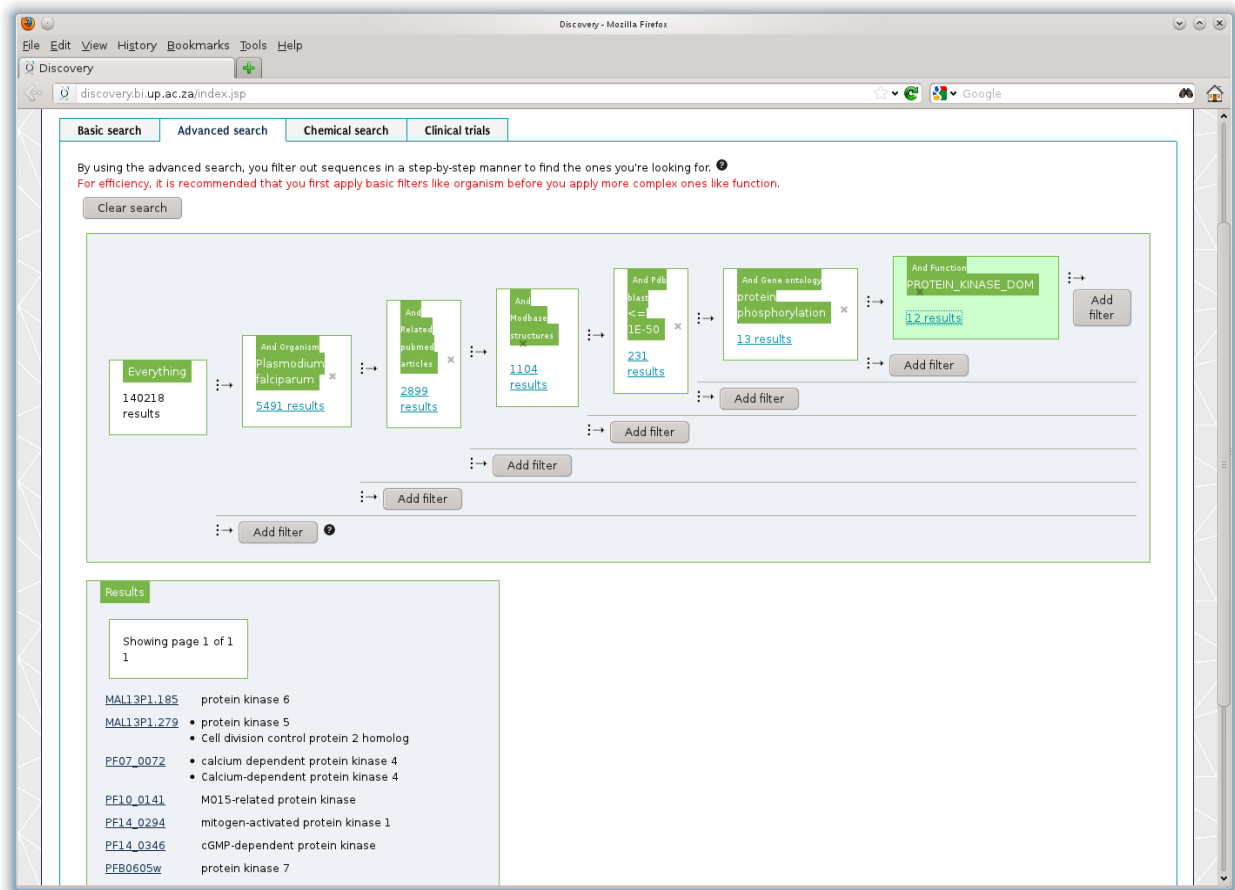


Figure 3.15: An advanced search in Discovery 2.0 for identifying a protein sequence belonging to the protein kinase superfamily in *P. falciparum*.

tion” is a biological process term with an accession “GO:0006468”. With this information, a *gene ontology filter*, using the search term “protein phosphorylation”, was further added to the 231 protein sequences. This resulted in 13 protein sequences which met all required criteria thus far and involved in protein phosphorylation.

The InterPro database (<http://www.ebi.ac.uk/interpro/>) was searched with the terms “protein kinase catalytic domain” to identify the domain signatures/descriptions which could be used in the *functions filter* to exclude all proteins without the protein kinase catalytic domain. The results showed that the domain “protein kinase, catalytic domain” (IPR000179) has contributing signatures PF00069 (Pkinase) and PS50011 (PROTEIN_KINASE_DOM). The search term “PROTEIN_KINASE_DOM” was added to the *functions filter* to identify all proteins with a protein kinase catalytic domain from 13 protein sequences (Figure 3.15). This resulted in only 12 proteins, and the protein “MO15-related protein kinase” (PlasmoDB identifier PF10_0141) was selected for analysis (Table 3.2).

Table 3.2: Summary of the annotation data for MO15-related protein kinase (PF10_0141) from *P. falciparum*.

Category	Type of annotation	Annotation
Summary	Names	- MO15-related protein kinase
	Sequence length	324 amino acids
	Protein Identifiers	- PF10_0141 (PlasmoDB) - Q8IJQ1 (UniProt) - pfa:PF10_0141 (KEGG GENE)
	PubMed articles	- 1
Function	Families	- None
	Domains	- Protein kinase, catalytic domain (IPR000719) - Serine/threonine- / dual-specificity protein kinase, catalytic domain (IPR002290) - Protein kinase-like domain (IPR011009)
	Sites	- Tyrosine-protein kinase, active site (IPR008266) - Protein kinase, ATP binding site (IPR017441)
Gene ontology	Cellular components	- GO:0005575 (cellular_component)
	Molecular functions	- GO:0000166 (nucleotide binding) - GO:0004672 (protein kinase activity) - GO:0004693 (cyclin-dependent protein kinase activity) - GO:0004713 (protein tyrosine kinase activity) - GO:0005524 (ATP binding) - GO:0016772 (transferase activity, transferring phosphorus-containing groups)
	Biological process	- GO:0006468 - (protein phosphorylation) - GO:0007049 - (cell cycle) - GO:0051726 - (regulation of cell cycle)
Orthology	<i>H. sapiens</i> orthologs	- None
	<i>A. gambiae</i> orthologs	- None
Structures	Top 3 PDB matches	- 1UA2; <i>E</i> -value: 6.31563e-64 - 3NIZ; <i>E</i> -value: 6.98477e-62 - 2QKR; <i>E</i> -value: 8.18468e-62
	MODBASE structures	- PF10_0141.1 (template 1CM8) - PF10_0141.2 (template 1UNL) - PF10_0141.3 (template 1UA2)
Metabolic pathways	KEGG	- Basal transcription factors - Nucleotide excision repair
	MPMP	- Protein kinase coding genes
	Reactome	- None
	EC numbers	- EC 2.7.11.22 (Cyclin-dependent kinase)
Interactions	DIP	- None
	MINT	- None
	IntAct	- None
Druggability	Top 3 domain matches	- 1V0B; <i>E</i> -value: 1.70962e-60; (druggable) - 1UA2; <i>E</i> -value: 2.84491e-59; (druggable) - 1V0P; <i>E</i> -value: 6.49845e-59; (druggable)
	Top 3 gene matches	- P50613; <i>E</i> -value: 1.06017e-64; (druggable) - Q5CRJ8; <i>E</i> -value: 6.8187e-62; (druggable) - Q07785; <i>E</i> -value: 6.27468e-61; (druggable)

The annotation data in Table 3.2 reveals that MO15-related protein kinase has all three protein identifiers from PlasmoDB, UniProt and KEGG GENE. Its protein sequence is 324 amino acids long and has only one article associated with it in Discovery 2.0. The InterProScan analysis matched three domains and two sites to the protein sequence, but no family entry associated with the sequence. GO terms were found for cellular component, molecular function and biological process. The information we can gather from the GO terms is that this protein is involved in regulation of the cell cycle, protein phosphorylation and has protein kinase activity. The protein kinase also has no ortholog in *H. sapiens* nor *A. gambiae*.

Looking at the structural information, the top three PDB-BLAST hits were crystal structures of the cyclin-dependent protein kinases (CDK) from *H. sapiens* (1UA2) and *Cryptosporidium parvum* (3NIZ and 2QKR). This information means that the *P. falciparum* MO15-related protein kinase has no experimentally solved crystal structure. However, the protein has three predicted structures from the MODBASE database. The protein is involved in two pathways from KEGG PATHWAY (“basal transcription factors” and “nucleotide excision repair”) and “protein kinase coding genes” pathway from MPMP. No pathway for this protein was identified in Reactome. The information gathered from analyzing these pathways is that the *P. falciparum* MO15-related protein kinase (EC 2.7.11.22) is a transcription factor that forms part of the Holo-FTHIH complex involved in nucleotide excision and repair. No protein-protein interaction data was observed for this protein. The druggability data reveals that the top three domains and the top three genes from DrugEBility that matched this protein were all druggable.

3.4.2 G protein-coupled receptor

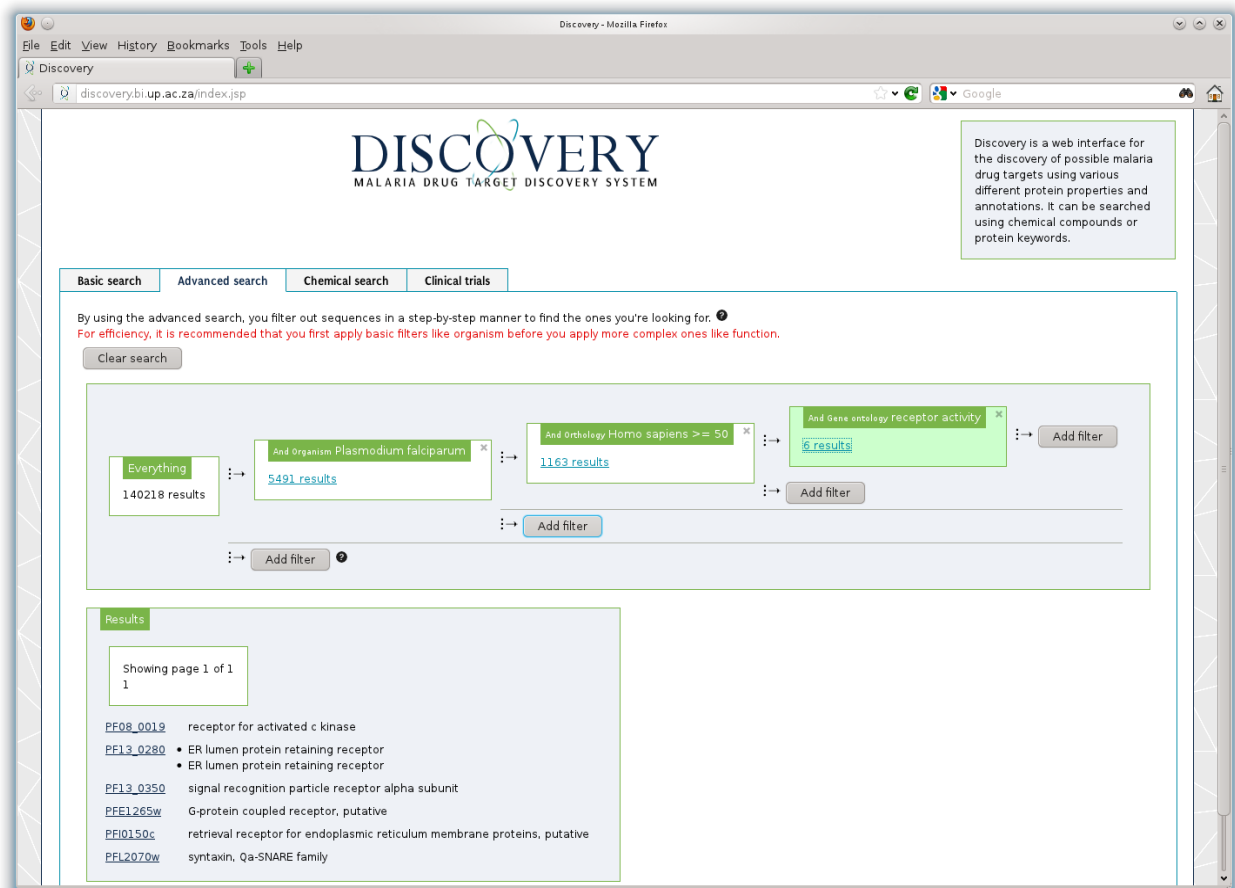
The G protein-coupled receptors (GPCR) superfamily is one of the most diverse families of membrane-bound receptors. They are responsible for recognition of intracellular messenger molecules and sensory messages that allow for communication of cells with each other and also with the environment (Bockaert and Pin, 1999; Wess, 1997). This is achieved by controlling the activity of enzymes, ion channels and transport of vesicles through interaction with specific hetero-trimeric G-proteins (consisting of α -, β - and γ - subunits) (Wess, 1997). GPCRs can recognize and transduce diverse messages such as light, hormones, odorants, nucleotides, lipids

and proteins. GPCRs are found in almost all eukaryotic organisms, however, comparison of different sequences reveal no sequence similarity between the different GPCR receptor families (Bockaert and Pin, 1999; Fredriksson and Schiöth, 2005).

Despite the lack of sequence similarity between the different GPCR families, GPCRs share a common central domain. This common structural feature constitutes of seven transmembrane α -helices (TM-1 to TM-VII) that are connected by three intracellular (i1 to i3) and three extracellular (e1 to e3) loops (Bockaert and Pin, 1999; Wess, 1997; Fredriksson and Schiöth, 2005). The extracellular region of this structure is involved in ligand binding, whilst the intracellular region is involved in G-protein recognition and activation (Wess, 1997). Mutations in GPCRs have been linked with many human diseases, and their ability to recognize a variety of small molecules has made GPCRs attractive drug targets. It is estimated that approximately 40% of drug targets are GPCRs (Hopkins and Groom, 2002; Fredriksson and Schiöth, 2005; Belmont *et al.*, 2006; Al-Lazikani *et al.*, 2008).

Identification of GPCRs in the species that are involved in malaria could perhaps be a step towards finding a cure for malaria. The method of identifying and classifying GPCRs in the genomes of *H. sapiens*, *M. musculus*, *T. rubripes*, *D. rerio*, *C. intestinalis*, *D. melanogaster*, *A. gambiae*, *C. elegans*, *O. sativa*, *A. thaliana*, *P. falciparum*, *S. pombe*, and *S. cerevisiae* by Fredriksson and Schiöth (2005) revealed that there were no GPCRs in *P. falciparum*. However, 865 and 268 genes were predicted to be GPCRs in *H. sapiens* and *A. gambiae*, respectively. The reason for no GPCR genes being identified in *P. falciparum* could perhaps be attributed to the high (A+T) content and longer gene length reported in *P. falciparum*, which limit sequence similarity searches (Gardner *et al.*, 2002).

Regardless of such information, an **advanced search** was carried out in Discovery 2.0 to identify a GPCR belonging to *P. falciparum*. The **organism filter** is again applied first since we are looking for a *P. falciparum* protein. Because we know that *H. sapiens* possess a larger number of GPCRs genes, the **orthology filter** was added to the search to find a *P. falciparum* ortholog in human (Figure 3.16). On the **orthology filter**, an organism “*H. sapiens*” was selected, and a cut-off value of “ $\geq 50\%$ change” was used to make the sequence similarity search less stringent, since sequence similarity searches with *P. falciparum* are limited by the (A+T) content and gene length. This reduced results from 5 491 to 1 163 protein sequences.



The screenshot shows the Discovery 2.0 web interface. At the top, there is a navigation bar with 'Basic search', 'Advanced search', 'Chemical search', and 'Clinical trials'. Below this, a search box contains the text 'Everything' with '140218 results'. A series of filters are applied: 'And organism Plasmodium falciparum' (5491 results), 'And orthology Homo sapiens >= 50' (1163 results), and 'And Gene ontology receptor activity' (6 results). The results section shows 'Showing page 1 of 1' and a list of protein sequences:

Protein ID	Description
PF08_0019	receptor for activated c kinase
PF13_0280	ER lumen protein retaining receptor
PF13_0350	signal recognition particle receptor alpha subunit
PFE1265w	G-protein coupled receptor, putative
PFI0150c	retrieval receptor for endoplasmic reticulum membrane proteins, putative
PFL2070w	syntaxin, Qa-SNARE family

Figure 3.16: An advanced search in Discovery 2.0 for identifying a GPCR protein sequence in *P. falciparum*.

To make the search more specific, the term “receptor activity” (a molecular function term with an accession “GO:0004872” when searched on the Gene Ontology database) was added to the *gene ontology filter*. This greatly reduced results from 1 163 to 6 protein sequences (Figure 3.16). The protein “G-protein coupled receptor, putative” (PlasmoDB identifier PFE1265w) was selected for analysis (Table 3.3).

This putative GPCR from *P. falciparum* is 467 amino acids long and is annotated with all three protein identifiers used in Discovery 2.0. Even though this is a putative protein, 35 PubMed articles were identified to be associated with this protein (Table 3.3). With so many articles, one would conclude that a lot of studies have been done on this protein. Analysis of these articles reveal that none of these articles are associated with the GPCR from *P. falciparum*. These articles, however, serve as a starting point for users interested in knowing about the protein. The InterProScan matched one family entry to this protein, IPR007822 (Lanthionine synthetase C-like). One would expect to see a GPCR fam-

Table 3.3: Summary of the annotation data for the putative G-protein coupled receptor (PFE1265w) from *P. falciparum*.

Category	Type of annotation	Annotation
Summary	Names	- G-protein coupled receptor, putative
	Sequence length	467 amino acids
	Protein Identifiers	- PFE1265w (PlasmoDB) - Q8I3L1 (UniProt) - pfa:PFE1265w (KEGG GENE)
	PubMed articles	- 35
Function	Families	- Lanthionine synthetase C-like (IPR007822)
	Domains	- None
	Sites	- None
Gene ontology	Cellular components	- GO:0005887 (integral to plasma membrane)
	Molecular functions	- GO:0003824 (catalytic activity) - GO:0004872 (receptor activity)
	Biological processes	- GO:0008152 (metabolic process)
Orthology	<i>H. sapiens</i> orthologs	- ENSP00000233714 (77.63% change) - ENSP00000254770 (78.77% change) - ENSP00000388713 (77.63% change) - ENSP00000393323 (77.63% change) - ENSP00000393597 (77.63% change) - ENSP00000395442 (97.18% change) - ENSP00000396518 (90.60% change) - ENSP00000397646 (77.63% change)
	<i>A. gambiae</i> orthologs	- None
Structures	Top 3 PDB matches	- 3E6U; <i>E</i> -value: 2.4241e-32
	MODBASE structures	- PFE1265w.1 (template 2G0D) - PFE1265w.2 (template 2G0D)
Metabolic pathways	KEGG	- None
	MPMP	- Established and putative Maurers clefts proteins
	Reactome	- None
	EC numbers	- None
Interactions	DIP	- None
	MINT	- None
	IntAct	- None
Druggability	Top 3 domain matches	- 3E73; <i>E</i> -value: 2.1819e-31; (not druggable) - 3E73; <i>E</i> -value: 2.1819e-31; (druggable) - 3E6U; <i>E</i> -value: 2.1819e-31; (not druggable)
	Top 3 gene matches	- O43813; <i>E</i> -value: 2.11e-32; (not druggable)

ily or domain to be associated with this protein, however this was not the case. According to the information for the InterPro “Lanthionine synthetase C-like” family entry (<http://www.ebi.ac.uk/interpro/ISearch?query=IPR007822>), this superfamily is composed of a highly divergent group of peptide modifying enzymes.

A crystal structure of a protein (lantibiotic cyclase from *L. lactis*) belonging to this family was determined, and found to be a globular structure (outer ring of helices enclose an inner

ring of 7 shorter hydrophobic helices). The 7 hydrophobic helical inner ring structure led to authors in classifying various proteins members of the lanthionine synthetase C-like family to belong to the GPCRs as this structure is similar to the seven transmembrane α -helical structure of GPCRs. GO terms from all three domains were identified for this putative protein and eight human orthologs (most of which are “lanC-like proteins”) were identified. However, all the human orthologs have a %change of more than 70% with the putative GPCR from *P. falciparum*, meaning that they are significantly different.

A single PDB-BLAST hit with an E -value of $2.4241e-32$ was found for this protein (Table 3.3). This crystal structure belongs to a lanC-like protein 1 from *H. sapiens*. Two MODBASE structure, both modelled with a PDB template 2G0D, were identified for this putative protein. According to the metabolic pathway data from MPMP, the *P. falciparum* putative GPCR is associated with Maurer’s cleft, a membranous structure in the cytoplasm of infected human erythrocytes formed by the parasite which is used for protein sorting and export (Spycher *et al.*, 2006). The protein is not annotated with any EC number and no metabolic pathway information was obtained from neither Reactome nor KEGG. The protein-protein interaction data was also not observed. Of the top three domains from DrugEBIity that matched with the protein, only one domain was druggable. The only gene from DrugEBIity matching the protein sequence was found not to be druggable.

3.4.3 Peptidase

Peptidases are enzymes responsible for the breaking down of proteins and polypeptides through cleavage of peptide bonds. They are also known as proteases, proteinases or peptide hydrolases, and are involved in many vital intracellular and extracellular processes, including the life cycle of many disease causing pathogens and viruses (Mittl and Grütter, 2006; Atkinson *et al.*, 2009). Peptidases are classified into groups based on the active site residue that provides the nucleophilic attack during the hydrolysis of protein/polypeptides. These groups are: cysteine, serine, aspartic, metallo and threonine peptidases (Coombs and Mottram, 1997; Sajid and McKerrow, 2002; Mittl and Grütter, 2006; Atkinson *et al.*, 2009).

Peptidases are considered good drug targets for many diseases caused by parasites. This is because for successful infection, parasitic organisms rely on their peptidases to pass through

the tissue and cellular barriers, degrade host proteins for nutrition, evade the immune system and also to process its own proteins (Atkinson *et al.*, 2009). For malaria, a group of aspartic peptidases from *P. falciparum* have been shown to be essential for hemoglobin degradation in the intraerythrocytic stages of the parasite, which provide nutrients for their own growth (Coombs *et al.*, 2001; Dash *et al.*, 2003; Prade *et al.*, 2005; Mittl and Grütter, 2006). These aspartic proteases, called plasmepsins (PMI, PMII and PMIV), are found in the food vacuole of the parasite and have been identified as attractive malaria drug targets since inhibition causes parasite death, suggesting that these enzyme play an essential role on the survival of the parasite.

To identify these enzymes belonging to *P. falciparum* in Discovery 2.0 using the **advanced search**, the **organism filter** was first added and “*P. falciparum*” was selected (Figure 3.17). This was followed by adding the term “proteolysis” (a biological process term with an accession “GO:0006508” when searched on the Gene Ontology database) on the **gene ontology**

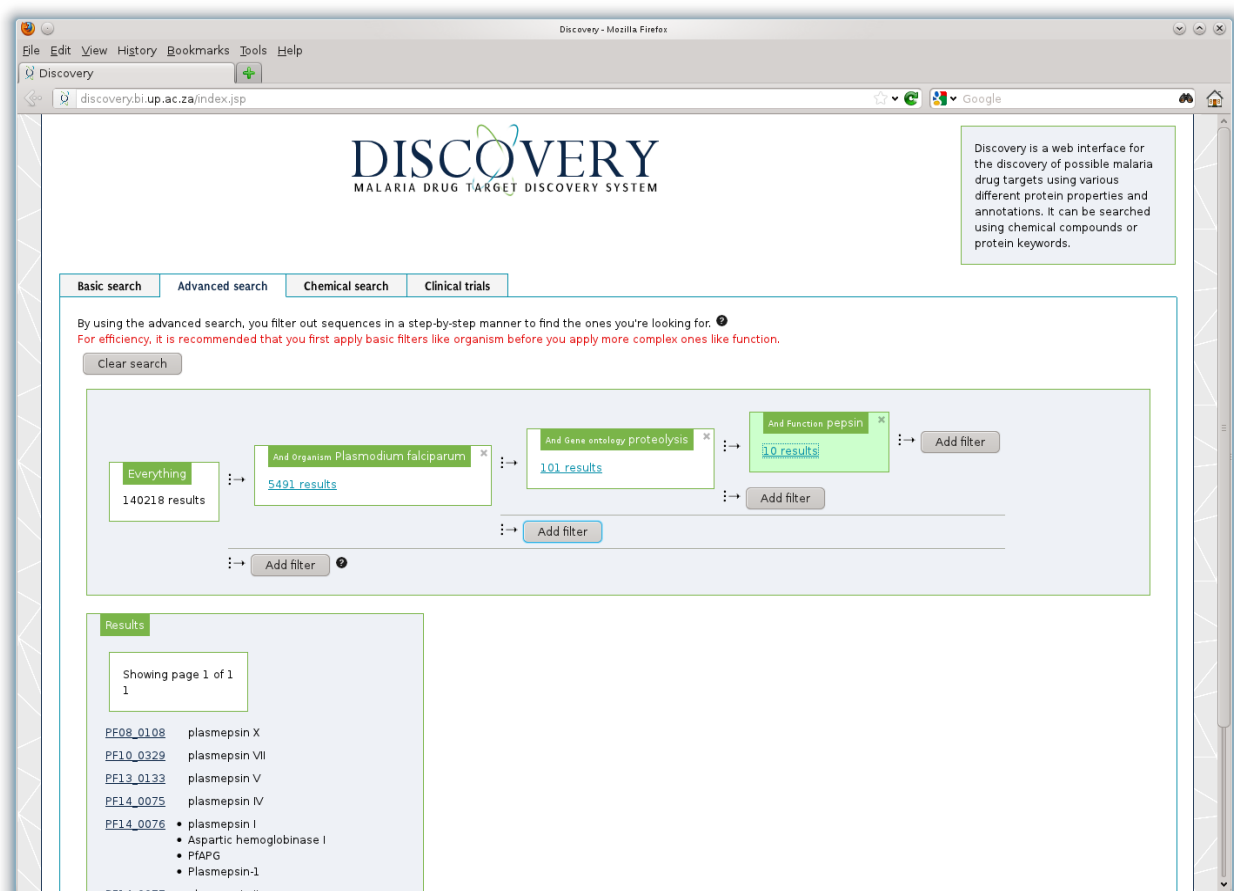


Figure 3.17: An advanced search carried out in Discovery 2.0 to identify aspartic proteases sequences in *P. falciparum*.

filter, which reduced results from 5 491 to 101 protein sequences. According to the MEROPS database (<http://merops.sanger.ac.uk/>) for peptidases, plasmepsins belong to the A1 family of aspartic proteases (Rawlings *et al.*, 2012). Searching the InterPro database with the term “A1 peptidase family”, a family hit “Peptidase A1” (IPR001461) is returned with contributing signatures PF00026 (Asp), PR00792 (PEPSIN), and PTHR13683 (PTHR13683). The search term “PEPSIN” was added to the **functions filter** to find protein sequences only belonging to the A1 family of aspartic proteases in *P. falciparum* (Figure 3.17). Only 10 plasmepsin protein sequences (plasmepsin I to X) were returned with this filter, and “plasmepsin I” (PlasmoDB identifier PF14_0076) was selected for analysis (Table 3.4).

The *P. falciparum* plasmepsin I, also known as aspartic hemoglobinase I, was annotated with all protein identifiers used in Discovery 2.0 and found to be associated with 111 PubMed articles. Its protein sequence is 452 amino acids long. The InterProScan program matched one family, two domains and one site entries with the protein sequence. All three GO domains were annotated with GO terms. According to the GO data, we may conclude that the plasmepsin I is located in the food vacuole of the *P. falciparum* parasite and involved in the catabolism of hemoglobin as mentioned before. Four *A. gambiae* proteins and seven *H. sapiens* proteins were found to be orthologous to plasmepsin I (Table 3.4). However, these proteins are significantly different to plasmepsin I as they have a %change of more than 70%. A crystal structure for plasmepsin I (3QRV) has been experimentally determined as shown by the top three BLAST-PDB results. The two other PDB-BLAST hits (2BJU and 1PFZ) were crystal structures of *P. falciparum* plasmepsin II.

Plasmepsin I is annotated with the EC number 3.4.23.38 and involved in the metabolic pathway “hemoglobin digestion and ferriprotoporphyrin IX polymerization” according to MPMP. The data from MPMP also shows that plasmepsin I is part of the “peptidases and proteases”, “S-glutathionylated proteins” and “nuclear genes with apicoplast signal sequences”. No protein-protein interaction data was observed for this protein. The top three DrugEBility domains and gene matching the protein were all druggable. However, even though the protein has an experimentally determined crystal structure, its druggability calculations were not found in the DrugEBility database (Table 3.4).

Table 3.4: Summary of the annotation data for plasmepsin I (PF14_0076) from *P. falciparum*.

Category	Type of annotation	Annotation
Summary	Names	- plasmepsin I - Aspartic hemoglobinase I - PfAPG - Plasmepsin-1
	Sequence length	452 amino acids
	Protein Identifiers	- PF14_0076 (PlasmoDB) - Q7KQM4 (UniProt) - pfa:PF14_0076 (KEGG GENE)
	PubMed articles	- 111
Function	Families	- Peptidase A1 (IPR001461)
	Domains	- Peptidase aspartic (IPR021109) - Peptidase aspartic, catalytic (IPR009007)
	Sites	- Peptidase aspartic, active site (IPR001969)
Gene ontology	Cellular components	- GO:0005773 (vacuole) - GO:0020020 (food vacuole)
	Molecular functions	- GO:0004190 (aspartic-type endopeptidase activity) - GO:0008233 (peptidase activity) - GO:0016787 (hydrolase activity)
	Biological processes	- GO:0006508 (proteolysis) - GO:0042540 (hemoglobin catabolic process)
Orthology	<i>H. sapiens</i> orthologs	- ENSP00000236671 (79.67% change); - ENSP00000272190 (79.00% change) - ENSP00000356163 (79.17% change); - ENSP00000356164 (89.67% change) - ENSP00000404902 (86.50% change); - ENSP00000415036 (85.33% change) - ENSP00000415840 (80.31% change)
	<i>A. gambiae</i> orthologs	- AGAP003277-PA (80.17% change) - AGAP003277-PB (80.17% change) - AGAP003277-PC (80.17% change) - AGAP003277-PD (80.17% change)
Structures	Top 3 PDB matches	- 3QRV; <i>E</i> -value: 0.0 - 2BJU; <i>E</i> -value: 0.0 - 1PFZ; <i>E</i> -value: 9.55036e-166
	MODBASE structures	- PF14_0076.1 (template 1MIQ) - PF14_0076.2 (template 1B5F)
Metabolic pathways	KEGG	- None
	MPMP	- Nuclear genes with apicoplast signal sequences - S-Glutathionylated proteins - Hemoglobin digestion and ferriprotoporphyrin IX polymerization - Peptidases and proteases
	Reactome	- None
	EC numbers	- EC 3.4.23.38 (Plasmepsin I.)
Interactions	DIP	- None
	MINT	- None
	Int Act	- None
Druggability	Top 3 domain matches	- 1M43; <i>E</i> -value: 6.61235e-147; (druggable) - 1M43; <i>E</i> -value: 6.61235e-147; (druggable) - 1LEE; <i>E</i> -value: 6.61235e-147; (druggable)
	Top 3 gene matches	- P46925; <i>E</i> -value: 0.0; (druggable) - O60990; <i>E</i> -value: 3.5948e-180; (druggable) - O60989; <i>E</i> -value: 8.36165e-175; (druggable)

3.4.4 Aminopeptidase

The advanced search in Discovery 2.0 does not only allow users to filter out protein sequences in a step-by-step manner, but it also allows users to create a work-flow where results may be split into two or more branches, allowing users to apply different filters on a set of results. To demonstrate this feature, a search was carried out to identify a peptidase in *P. falciparum* that has an experimentally determined crystal structure and share less sequence similarity with *H. sapiens*. On the previous search done to identify plasmepsins in *P. falciparum* (Figure 3.17), a branch on the results after adding the *gene ontology filter* was added. A *PDB-BLAST filter*, with an *E*-value cut-off of “ ≤ 0 ”, was added to find peptidases with experimentally solved crystal structures (Figure 3.18).

Adding this filter reduced the results from 101 to only 7 protein sequences. An *orthology*

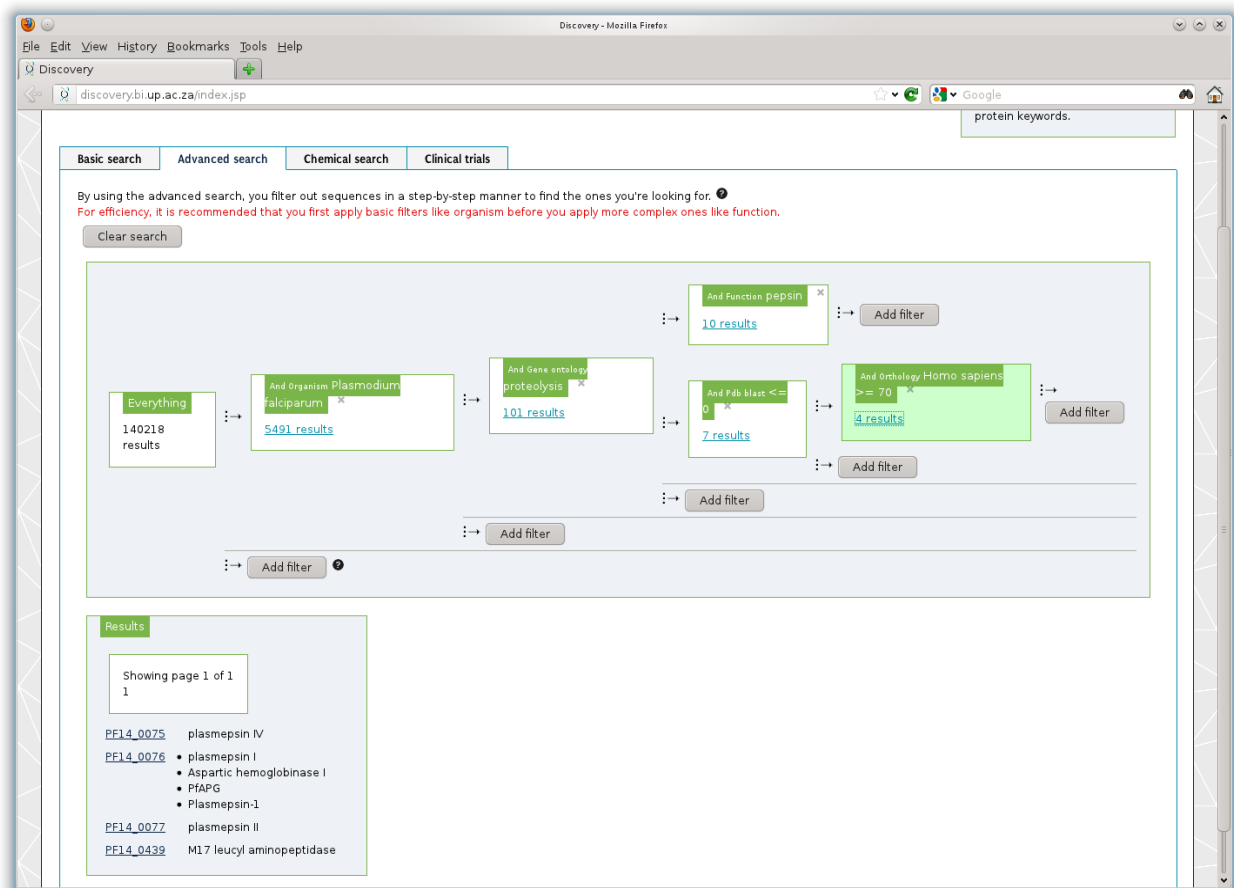


Figure 3.18: An advanced search in Discovery 2.0 for identifying a peptidase in *P. falciparum*. The search was carried out to find a peptidase in *P. falciparum* with an experimentally solved crystal structure as well as no significant similarity to *H. sapiens* proteins. The search also demonstrates the branching of results, where one or more filters may be applied to the same results.

filter was then added to exclude protein sequences that share more than 30% sequence similarity (%change of more than 70%). *H. sapiens* was selected on the *orthology filter* and a

Table 3.5: Summary of the annotation data for M17 leucyl aminopeptidase (PF14_0439) from *P. falciparum*.

Category	Type of annotation	Annotation
Summary	Names	- M17 leucyl aminopeptidase
	Sequence length	605 amino acids
	Protein Identifiers	- PF14_0439 (PlasmoDB) - Q8IL11 (UniProt) - pfa:PF14_0439 (KEGG GENE)
	PubMed articles	- 1
Function	Families	- Peptidase M17 (IPR011356)
	Domains	- Peptidase M17, leucyl aminopeptidase, C-terminal (IPR000819)
	Sites	- None
Gene ontology	Cellular components	- GO:0005622 (intracellular) - GO:0005737 (cytoplasm)
	Molecular functions	- GO:0004177 (aminopeptidase activity) - GO:0008235 (metalloexopeptidase activity) - GO:0030145 (manganese ion binding) - GO:0046872 (metal ion binding)
	Biological processes	- GO:0006508 (proteolysis) - GO:0019538 (protein metabolic process)
Orthology	<i>H. sapiens</i> orthologs	- ENSP00000226299 (77.97% change)
	<i>A. gambiae</i> orthologs	- AGAP003869-PA (77.97% change)
Structures	Top 3 PDB matches	- 3KQX; <i>E</i> -value: 0.0 - 3H8E; <i>E</i> -value: 1.30724e-63 - 1GYT; <i>E</i> -value: 5.40761e-58
	MODBASE structures	- PF14_0439.1 (template 1GYT) - PF14_0439.2 (template 1GYT)
Metabolic pathways	KEGG	- Glutathione metabolism - Metabolic pathways
	MPMP	- Nuclear genes with apicoplast signal sequences - S-Glutathionylated proteins - Hemoglobin digestion and ferriprotoporphyrin IX polymerization - Peptidases and proteases
	Reactome	- None
	EC numbers	- 3.4.11.1 (Leucyl aminopeptidase)
Interactions	DIP	- Q8IL11 (direct interaction)
	MINT	- None
	IntAct	- None
Druggability	Top 3 domain matches	- 3KR4; <i>E</i> -value: 1.06665e-173; (druggable) - 3KR4; <i>E</i> -value: 1.06665e-173; (druggable) - 3KQX; <i>E</i> -value: 1.06665e-173; (druggable)
	Top 3 gene matches	- Q8IL11; <i>E</i> -value: 0.0; (not druggable) - O86436; <i>E</i> -value: 1.1658e-63; (not druggable) - P68767; <i>E</i> -value: 4.8225e-58; (druggable)

value of “ ≥ 70 ” % change was added, which reduced the results to only 4 protein sequences, of which three were plasmepsins and one was an aminopeptidase (M17 leucyl aminopeptidase) (Figure 3.18). M17 leucyl aminopeptidase (PF14_0439) was selected for analysis (Table 3.5).

The results for M17 leucyl aminopeptidase in Discovery 2.0 show that the protein is associated with only one PubMed article. The protein also has all three protein identifiers PlasmDB, UniProt and KEGG GENE. Its protein sequence is 605 amino acids long. Only one family and one domain entries from InterPro were matched with the protein; no conserved sites were found with the InterProScan program. From the GO data (Table 3.5), we may conclude that the protein is responsible for protein degradation in the cytoplasm. One hypothetical protein from *A. gambiae* and a leucine aminopeptidase 3 (LAP-3) from *H. sapiens* were found to be orthologs of the *P. falciparum* M17 leucyl aminopeptidase, both having a %change of more than 70%.

The PDB-BLAST results show that the protein does have an experimentally determined crystal structure (3KQX). Two MODBASE structures have been predicted for the protein using the PDB structure 1GYT as a template. The metabolic pathway data shows that this protein is associated with the same pathways in MPMP as plasmepsin I (Table 3.4 and Table 3.5) and has an EC number 3.4.11.1. Two additional pathways from KEGG PATHWAYS, “glutathione metabolism” and “metabolic pathways”, were associated with M17 leucyl aminopeptidase. No pathways from Reactome were found. Analyzing the protein-protein interaction data, it is found that the protein has a direct interacts with itself (interaction data from DIP). The top three domains from DrugEBility matching the protein were all druggable (Table 3.5). The protein also has its druggability data calculated since it has its own structure as shown by the top three gene matches from DrugEBility (Q8IL11). However, the protein was found not to be druggable.

3.4.5 Dehydrogenase

In Section 1.3.2, the protein *PfDHOD* was used to demonstrate how important it is to have assay information for a potential drug target in order to facilitate the drug discovery process. A number of HTS screens as well as *in silico* dockings of small molecules active against *PfDHOD* have been described (Baldwin *et al.*, 2005; Patel *et al.*, 2008). To evaluate the annotation data for this protein in Discovery 2.0, an advanced search was carried to identify the protein using

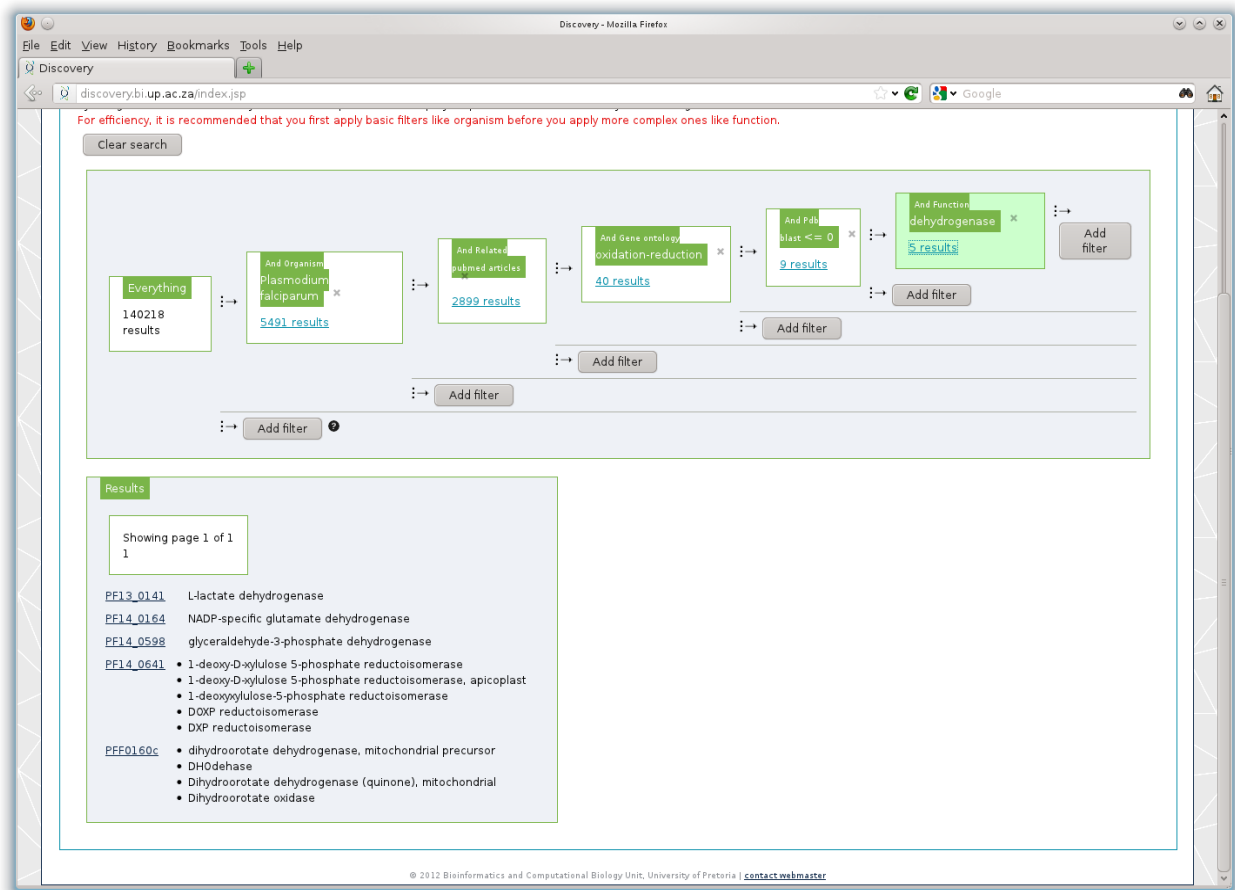


Figure 3.19: An advanced search in Discovery 2.0 carried out to identify the *P. falciparum* enzyme DHOD (*PfDHOD*).

the the information mentioned in this study to filter out unwanted protein sequences. First, we know that the enzyme belongs to *P. falciparum* and that a number of articles have been published on it, thus we can apply *organism filter* first followed by the *related PubMed articles filter* to exclude unwanted protein sequences.

Secondly, we know that the protein is involved in a reduction-oxidation (redox) reaction, and this can be applied to the *gene ontology filter* (“reduction-oxidation” is a biological process term with an accession “GO:0055114” on the Gene Ontology database). We also know that *in silico* docking studies have been done for *PfDHOD*; and because *in silico* dockings require a crystal structure of a protein, we can assume that the protein has a solved crystal structure and use this information on the *PDB-BLAST filter*. Lastly, *PfDHOD* is a dehydrogenase enzyme, and this information can be used in the *functions filter*.

Table 3.6: Summary of the annotation data for *PfDHOD* (PFF0160c) from *P. falciparum*.

Category	Type of annotation	Annotation
Summary	Names	- dihydroorotate dehydrogenase, mitochondrial precursor - DHO dease - Dihydroorotate dehydrogenase (quinone), mitochondrial - Dihydroorotate oxidase
	Sequence length	569 amino acids
	Protein Identifiers	- PFF0160c (PlasmoDB) - Q08210 (UniProt) - pfa:PFF0160c (KEGG GENE)
	PubMed articles	- 54
Function	Families	- Dihydroorotate dehydrogenase, class 1/ 2 (IPR012135) - Dihydroorotate dehydrogenase, class 2 (IPR005719)
	Domains	- Aldolase-type TIM barrel (IPR013785)
	Sites	- Dihydroorotate dehydrogenase, conserved site (IPR001295)
Gene ontology	Cellular components	- GO:0005739 (mitochondrion) - GO:0005743 (mitochondrial inner membrane) - GO:0016020 (membrane) - GO:0016021 (integral to membrane)
	Molecular functions	- GO:0003824 (catalytic activity) - GO:0004152 (dihydroorotate dehydrogenase activity) - GO:0016491 (oxidoreductase activity)
	Biological processes	- GO:0006207 ('de novo' pyrimidine nucleobase biosynthetic process) - GO:0006221 (pyrimidine nucleotide biosynthetic process) - GO:0006222 (UMP biosynthetic process) - GO:0044205 ('de novo' UMP biosynthetic process) - GO:0055114 (oxidation-reduction process)
Orthology	<i>H. sapiens</i> orthologs	- AGAP002037-PA (76.98% change)
	<i>A. gambiae</i> orthologs	- ENSP00000219240 (77.13% change)
Structures	Top 3 PDB matches	- 1TV5; <i>E</i> -value: 0.0 - 3I65; <i>E</i> -value: 0.0 - 3SFK; <i>E</i> -value: 0.0
	MODBASE structures	- PFF0160c.1 (template 1GTE) - PFF0160c.2 (template 1TV5)
Metabolic pathways	KEGG	- Pyrimidine metabolism - Metabolic pathways
	MPMP	- Mitochondrial electron flow - Nuclear genes with mitochondrial signal sequences - Pyrimidine metabolism
	Reactome	- Metabolism of nucleotides/Pyrimidine metabolism/Pyrimidine biosynthesis
	EC numbers	- EC 1.3.5.2 (Dihydroorotate dehydrogenase (quinone))
Interactions	DIP	- None
	MINT	- None
	IntAct	- None
Druggability	Top 3 domain matches	- 1TV5; <i>E</i> -value: 0.0; (not druggable) - 3O8A; <i>E</i> -value: 0.0; (not druggable) - 3KVK; <i>E</i> -value: 7.54887e-71; (not druggable)
	Top 3 gene matches	- Q08210; <i>E</i> -value: 0.0; (druggable) - Q63707; <i>E</i> -value: 1.50681e-73; (druggable) - Q02127; <i>E</i> -value: 9.45175e-71; (druggable)

Applying the *organism filter* (*P. falciparum*) followed by applying the *related PubMed articles filter*, the protein sequences were reduced down to 2 899 (Figure 3.19). Adding the “oxidation-reduction” search term on the *gene ontology filter* resulted in 40 protein sequences. A *PDB-BLAST filter* (*E*-value of 0) was further added, resulting in only 9 protein sequences, which were further reduced to 5 protein sequences by adding the search term “dehydrogenase” in the *functions filter* (Figure 3.19). *PfDHOD* (PFF0160c) was one of the 5 protein sequences and was selected for analysis (Table 3.6). In Discovery 2.0, *PfDHOD* is annotated with protein identifiers from PlasmoDB, UniProt and KEGG GENE. Its protein sequence is 569 amino acids long and has 54 related PubMed articles. InterPro entries were found for family, domain and functional sites. The information that can be gathered from the GO data is that it found in the inner membrane of the mitochondria and involved in the biosynthesis of pyrimidine nucleotides (Table 3.6).

PfDHOD has one ortholog in *A. gambiae* one ortholog in *H. sapiens*, both having a %change of more than 70%. All top three PDB-BLAST results are of different crystal structures of *PfDHOD* having an *E*-value of 0. Two MODBASE structures were predicted for *PfDHOD* using PDB structures 1GTE and 1TV5. The data from MPMP shows that *PfDHOD* (EC:1.3.5.2) is involved in the “mitochondrial electron flow” and “pyrimidine metabolism” pathways. The protein is also part of the “nuclear genes with mitochondrial signal sequences”. KEGG PATHWAYS data shows that the protein is associated with “pyrimidine metabolism” and “metabolic pathways”. In Reactome, the protein is associated with “metabolism of nucleotides”. No protein-protein interaction data was found for *PfDHOD*. The top three domains from DrugEBility that matched *PfDHOD* were not druggable. However, the top three genes matching the protein (including the *PfDHOD* gene) were all druggable.

3.5 Assessment of a protein target using Discovery 2.0

The enzyme S-adenosyl-L-homocysteine hydrolase (SAHH) from *P. falciparum* (which will be referred to as *PfSAHH* from here on), involved in the cysteine and methionine metabolism, was chosen to demonstrate how the data available in Discovery 2.0 can be used to assess a protein on the six target assessment criteria mentioned in Section 1.3. This protein was selected as it has been proposed to be a possible malaria target (Shuto *et al.*, 2002). Since this is a

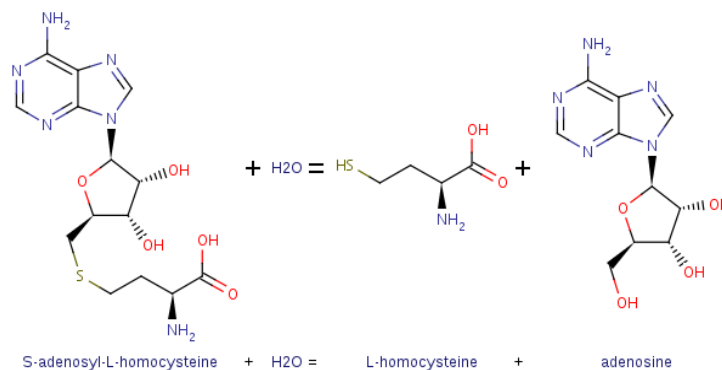


Figure 3.20: **Hydrolysis of S-adenosyl-L-homocysteine to adenosine and L-homocysteine.**

known potential malarial target, it is expected that much research has been done on it, and thus, it would be well annotated. *PfSAHH* is an enzyme that catalyzes the hydrolysis of S-adenosyl-L-homocysteine (SAH) to adenosine and L-homocysteine (Figure 3.20) (Creedon *et al.*, 1994; Shuto *et al.*, 2002). The breakdown of SAH is a reversible reduction/oxidation reaction involving the nicotinamide adenine dinucleotide (NAD) co-factor. To search for *PfSAHH* in Discovery 2.0, the user may enter either its PlasmoDB identifier (PFE1050w), its UniProt accession (P50250) or the name of the enzyme on the the **basic search**.

Summary

The **summary tab** shows that *PfSAHH* enzyme is also known as “adenosylhomocysteinase” or “AdoHcyase”. There are also links to PlasmoDB, UniProt and KEGG GENE. 50 PubMed articles were associated with this protein.

Function

The InterProScan results on the **function tab** for *PfSAHH* shows that three different types of InterPro identifiers were matched with the sequence i.e., IPR000043 (Adenosylhomocysteinase), IPR015878 (S-adenosyl-L-homocysteine hydrolase, NAD binding domain) and IPR020082 (S-adenosyl-L-homocysteine hydrolase, conserved site). The summary of the InterPro family, domain and site signatures matching to *PfSAHH* is shown in Table 3.7 along with the analysis methods used to match the signatures. Other signature that matched the *PfSAHH* protein sequence, but with no InterPro identifier, were G3DSA:3.40.50.1480 (Ad_hcy_hydrolase)

which was identified through the Gene3D, as well as SSF51735 (NAD(P)-bd) and SSF52283 (SSF52283) which were both identified by the superfamily analysis method.

Table 3.7: Summary of the InterPro signatures matching the *PfSAHH* protein sequence.

InterPro entry	Signatures	Analysis method
IPR000043 (Family)	- TIGR00936 (ahcY)	- HMMTigr
	- PF05221 (AdoHcyase)	- HMMPfam
	- PTHR23420 (Ad_hcy_hydrolase)	- HMMPanther
	- PIRSF001109 (Ad_hcy_hydrolase)	- HMMPiR
IPR015878 (Domain)	- PF00670 (AdoHcyase_NAD)	- HMMPfam
IPR020082 (Site)	- PS00738 (ADOHCYASE_1)	- PatternScan
	- PS00739 (ADOHCYASE_2)	- PatternScan

Gene ontology

The **gene ontology tab** shows that the GO terms at molecular function level associated with *PfSAHH* are GO:0004013 (adenosylhomocysteinase activity) and GO:0016787 (hydrolase activity), whilst GO:0006730 (one-carbon metabolic process) is associated with the protein at a biological process level. No GO term was associated with *PfSAHH* at cellular component level.

Orthology

The results from the OrthoMCL clustering on the **orthology tab** reveal that the SAHH is found in all eight species in Discovery 2.0. However, SAHH from *A. gambiae* and *H. sapiens* are shorter when compared to their *Plasmodium* counterparts. The T-coffee alignment of the cluster shows that there is a big gap in the *A. gambiae* and *H. sapiens* from amino acids 145 to 187 when compared to the *P. falciparum* and the other *Plasmodium* SAHH.

Structures

The PDB-BLAST results on the **structure tab** shows that the *PfSAHH* protein sequence has an exact match with a PDB entry 1V8B, which is an experimentally determined crystal structure of *PfSAHH*. This PDB hit has the highest score of 2 564 and lowest *E*-value of zero. Two MODBASE models have been predicted for the *PfSAHH* protein, PFE1050w.1 and PFE1050w.2 which were built on PDB templates 1LI4 and 1V8B respectively.

Metabolic pathways

On the **metabolic pathways tab**, the results show that the enzyme has an EC number 3.3.1.1. According to the MPMP database, the *PfSAHH* enzyme is associated with “methionine and polyamine metabolism”, “proteins targeted by the thioredoxin superfamily enzymes” (list of proteins but not a pathway) and “S-Glutathionylated proteins” (list of proteins but not a pathway). KEGG PATHWAY associated the enzyme with “cysteine and methionine metabolism” and “metabolic pathways” (non-specific global map of all metabolic pathways). The results from Reactome show that *PfSAHH* is involved in “metabolism of amino acids and derivatives” as well as “biological oxidations”.

Interactions

The results for protein-protein interactions on the **interactions tab** show that only the IntAct and MINT databases have interactions data for the *PfSAHH* enzyme. Both these databases show the same number of proteins, all from *P. falciparum*, to interact with *PfSAHH*. The UniProt accessions for these proteins that interact with the *PfSAHH* enzyme are Q8I561 (conserved protein, unknown function), Q8I2F7 (ring-exported protein [REX3]), Q8IFP1 (U5 small nuclear ribonucleoprotein-specific protein, putative), O96221 (Sec31p putative), Q8IKB6 (histone deacetylase, putative), Q8IIC8 (conserved protein, unknown function), Q8IBL5 (conserved protein, unknown function), Q8IAZ3 (Eukaryotic translation initiation factor 3 subunit G) and Q8IJY1 (conserved protein, unknown function). All identified interactions were through the two hybrid fragment pooling approach and the types of interactions between *PfSAHH* and the other proteins are physical associations.

Druggability

Analyzing the druggability of *PfSAHH* on the **druggability tab**, we see that the most significant BLAST domain matches (with the crystal structure of *PfSAHH*, 1V8B) are all not druggable. Analyzing the top three genes producing significant BLAST alignments with *PfSAHH* (P50250, P60176 and Q3JY79) also shows that these genes are not druggable. P50250 (the *PfSAHH* enzyme being discussed here) has an overall druggability of 0% whilst P60176 (SAHH from *Mycobacterium tuberculosis*) and Q3JY79 (SAHH from *Burkholderia pseudomallei*, strain

1710b) have a druggability of 48% and 45% respectively, as shown by the links to the DrugE-Bility database.

3.5.1 Essentiality

Essentiality assesses whether a protein is vital for the growth of the parasite or the progression of the disease in the human hosts and if inhibition of the protein can reverse the symptoms of the disease. The PlasmoDB, UniProt and KEGG gene links in the **summary tab** provides basic information about *PfSAHH*. Viewing the protein in PlasmoDB genomic browser (PlasmoDB GBrowse v2.39) reveals that the 479 amino acid long *PfSAHH* protein is encoded by a 1 440 nucleotide gene located on chromosome 5 (842 035 to 873 474) of the *P. falciparum* genome (Figure 3.21). *PfSAHH* has a molecular weight of 53 840 Da and an isoelectric point of 5.71. The predicted function data reveals that *PfSAHH* belongs to the adenosylhomocysteinase family of enzymes, has an NAD binding domain and two conserved sites.

The involvement of *PfSAHH* in the metabolism of methionine (Figure 3.22) makes it an

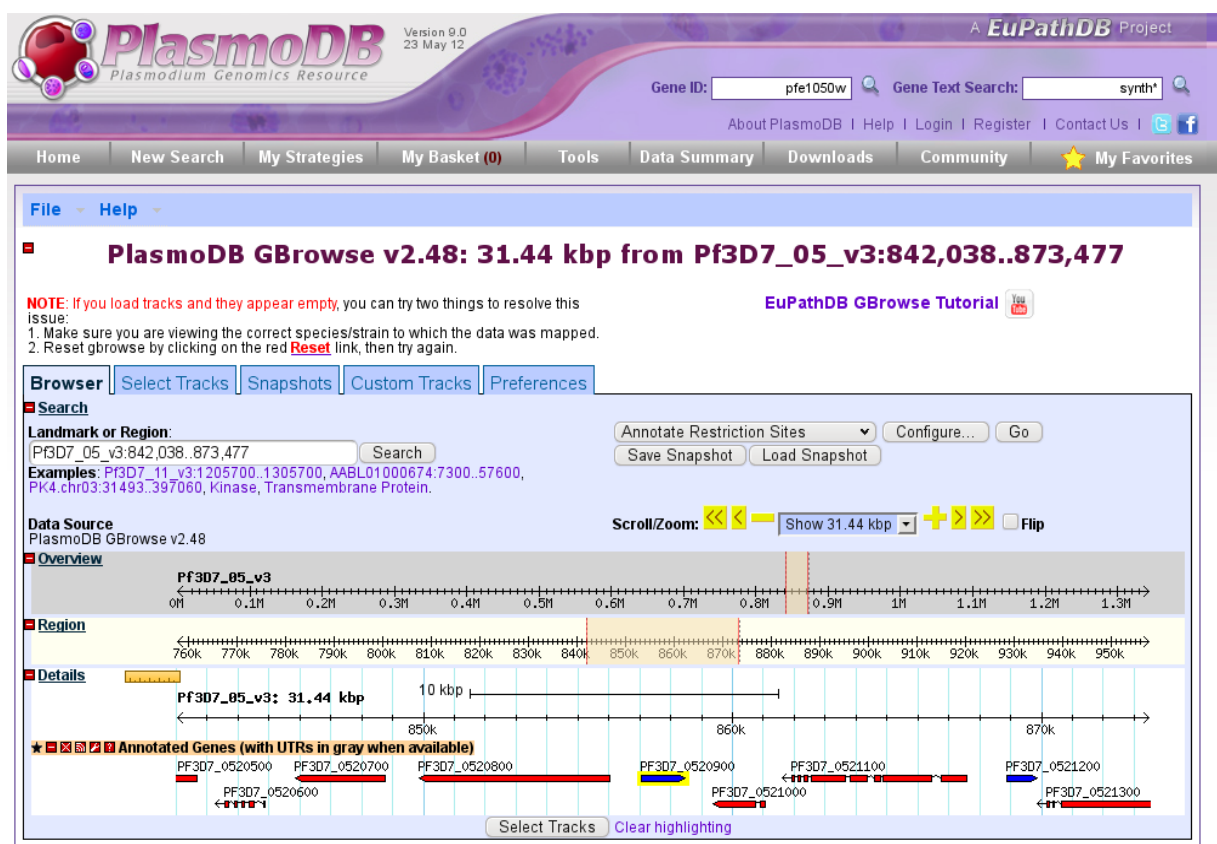


Figure 3.21: Analysis of *PfSAHH* in PlasmoDB genome browser.

Methionine and polyamine metabolism

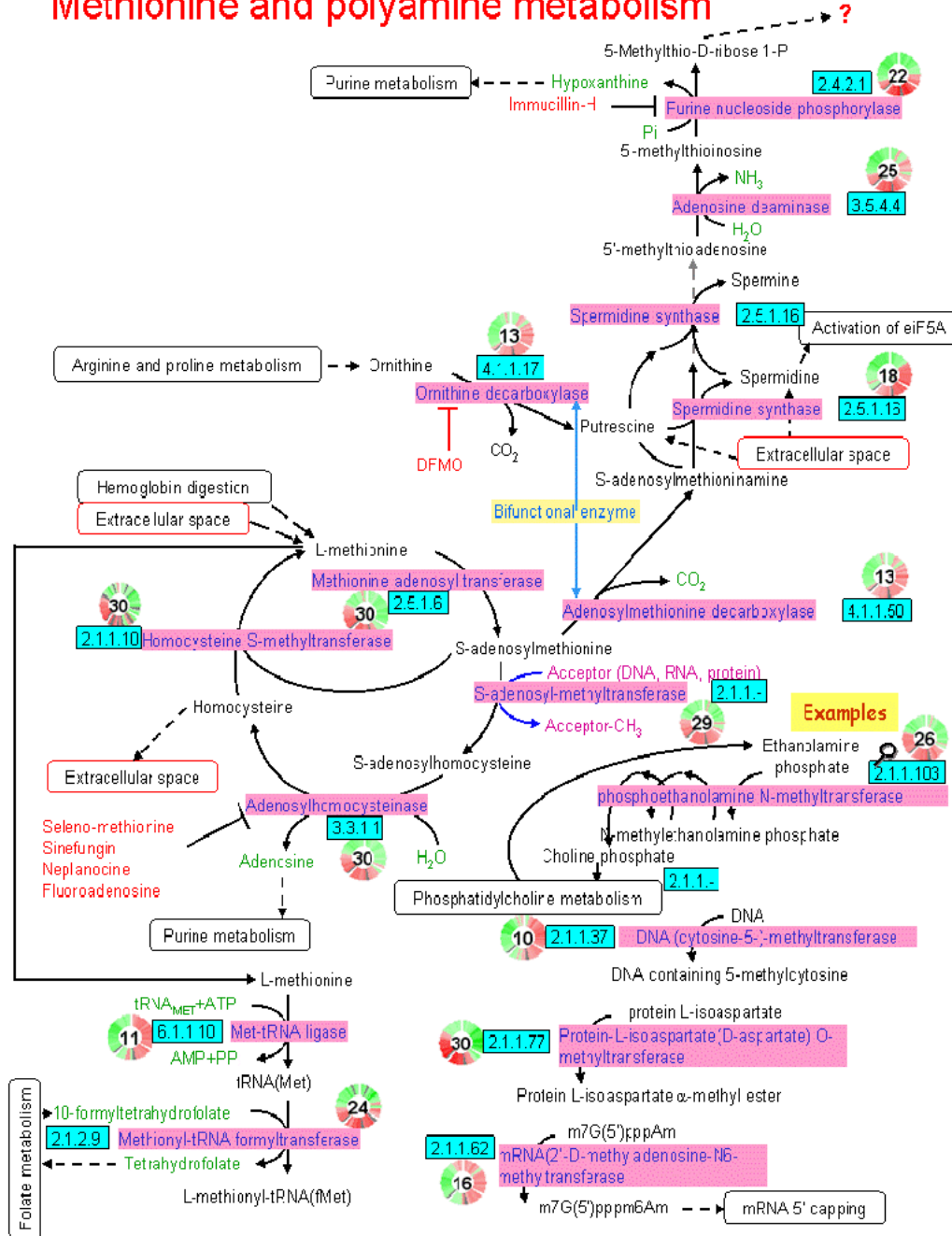


Figure 3.22: **Methionine and polyamine metabolism.** *PfSAHH* (EC 3.3.1.1) hydrolyzes SAH to adenosine and L-homocysteine. Adenosine is used in the purine metabolism pathway whilst L-homocysteine is (Adapted from <http://sites.huji.ac.il/malaria/maps/methioninemetpath.html>).

attractive target for malaria since it is a key regulator of the reactions involving methionine in the parasite. The *PfSAHH* enzyme (EC 3.3.1.1) hydrolyzes SAH (a product of S-adenosylmethionine (SAM)-dependent methyltransferases) to adenosine and L-homocysteine, a reaction that requires NAD as a co-factor (Figure 3.20). Inhibition of *PfSAHH* causes an accumulation of cellular SAH. Since SAH is both a product and an inhibitor of SAM-dependent

methyltransferases (which have an essential role in methylation of lipids, protein and nucleic acids), inhibiting the degradation of SAH results in a feedback inhibition of methylation reactions, thus disrupting of a number of metabolic pathways in the parasite (Creedon *et al.*, 1994; Nakanishi *et al.*, 2001; Shuto *et al.*, 2002; Tanaka *et al.*, 2004).

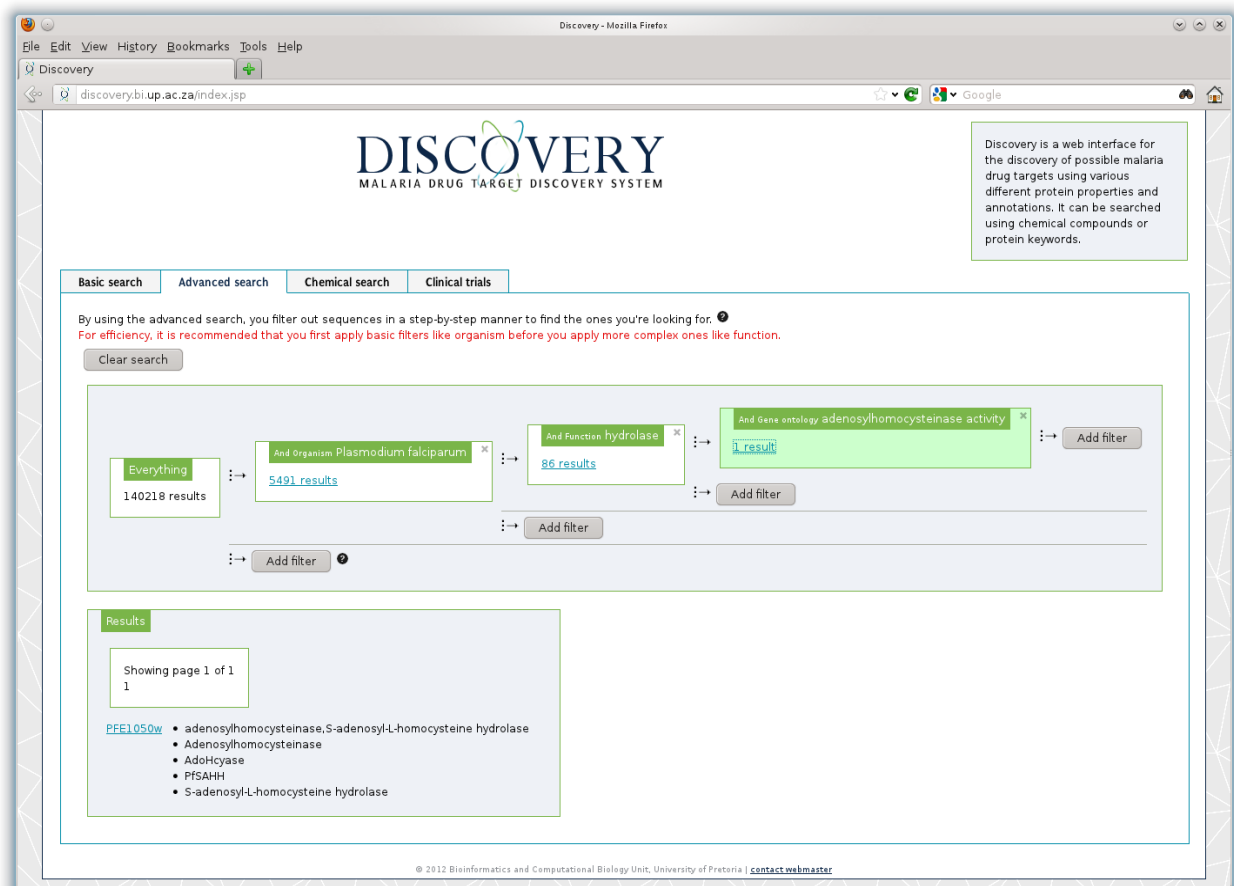
The transcriptomic clock for the *PfSAHH* enzyme in the methionine and polyamine metabolism pathway (Figure 3.22) reveals that the protein is not expressed in the early and late stages of the erythrocyte infection cycle. It is maximally expressed at 30 hours after the merozoites invade the erythrocytes. *PfSAHH* was also identified as a choke-point in the *in silico* metabolic network analysis, suggesting it plays an essential role in the survival of the *P. falciparum* in its host (Yeh *et al.*, 2004; Fatumo *et al.*, 2009; Huthmacher *et al.*, 2010). Literature evidence also supports this, as it was shown by Nakanishi *et al.* (2001) and Shuto *et al.* (2002) that the *PfSAHH* inhibitors were able to inhibit the growth of the parasites *in vivo* and *in vitro*, proving that *PfSAHH* is indeed essential for the survival of the malarial parasites.

3.5.2 Assay feasibility

With assay feasibility, the *PfSAHH* enzyme is assessed whether it can be readily expressed using available protocols and reagents, as well as to identify any described activity assays for *PfSAHH*. Following the link to BRENDA in the **metabolic pathways tab**, an article by Nakanishi *et al.* (2001) is found under the molecular properties of *PfSAHH*. This article describes the cloning, expression and purification of a recombinant *PfSAHH* protein in *E. coli* cells. The article also describes an activity assay for *PfSAHH*, which measures the formation of a nucleoside inosine by high pressure liquid chromatography (HPLC). In this assay, the *PfSAHH* enzyme is incubated with adenosine deaminase, an enzyme that converts adenosine to inosine, and the SAH substrate is added to start the reaction. The reaction is then stopped and the analyzed for the presence of inosine using HPLC (Nakanishi *et al.*, 2001). With such information at hand for *PfSAHH*, the drug discovery process can be greatly facilitated as the protein can be obtained in large quantities for HTS and an assays for the detection of activity is available.

3.5.3 Resistance

Resistance assesses the potential for the arousal of drug resistance for a particular protein chosen as a target. Here, the alternative pathways and isoforms that may replace the inhibited target are identified. The GO terms as well as ortholog information can be used to identify potential protein isoforms that may replace the function of an inhibited protein target, thus causing resistance to a particular drug. Genes that have the same function at a molecular level and found in the same compartment may be identified through the **advanced search** function in Discovery 2.0. By first adding an *organism filter* and selecting “*P. falciparum*”, then adding the *function filter* “hydrolase” (since *PfSAHH* is a hydrolase) and finally the *gene ontology filter* “adenosylhomocysteinase activity”, all the proteins with the same activity as *PfSAHH* were identified (Figure 3.23). Eighty-six proteins were returned which function as hydrolases but only one protein, PFE1050w (*PfSAHH*), was annotated with adenosylhomocysteinase activity with the above mentioned advanced search, meaning no other protein in *P. falciparum* can



The screenshot shows the Discovery 2.0 web interface in a Mozilla Firefox browser. The page title is "DISCOVERY MALARIA DRUG TARGET DISCOVERY SYSTEM". The search interface includes tabs for "Basic search", "Advanced search", "Chemical search", and "Clinical trials". The "Advanced search" tab is active, and a search query is displayed with the following filters: "Everything" (140218 results), "And Organism Plasmodium falciparum" (5491 results), "And Function hydrolase" (86 results), and "And Gene ontology adenosylhomocysteinase activity" (1 result). The "Results" section shows "Showing page 1 of 1" and lists the following results for PFE1050w:

- adenosylhomocysteinase,S-adenosyl-L-homocysteine hydrolase
- Adenosylhomocysteinase
- AdoHcyase
- PFSAHH
- S-adenosyl-L-homocysteine hydrolase

At the bottom of the page, there is a copyright notice: "© 2012 Bioinformatics and Computational Biology Unit, University of Pretoria | [contact webmaster](#)".

Figure 3.23: Advanced search to identify proteins with the same or similar function to *PfSAHH*.

replace the function of *PfSAHH* if it were inhibited. This information is also confirmed by the absence of any *PfSAHH* paralogs within *P. falciparum* as shown by the OrthoMCL results.

The *in silico* metabolic network analysis for identifying choke-points in *P. falciparum* pathways done by Yeh *et al.* (2004), and Huthmacher *et al.* (2010) also reveals that *PfSAHH* was identified as both a choke-point and a potential target in their analysis data. This shows that there are no possibilities of an alternative pathway that may replace the function of *PfSAHH*. This means *PfSAHH* is the only enzyme that can hydrolyze SAH to produce adenosine and L-homocysteine, and that it is a key regulator of methylation reactions in the parasite. Mutations in the genes may also pose as a major threat to drug resistance, especially if the mutations occur in the sites of proteins that allow for the binding of a drug or an inhibitor.

Tanaka *et al.* (2004) performed mutational analysis on the *PfSAHH* enzyme to determine the inhibitor sensitivity of the mutant *PfSAHH* enzyme. In this study, the Cys59 residue was examined by performing a single amino acid substitution to Thr59 (C59T). The inhibitory activities of the mutant *PfSAHH* harboring C59T mutation were tested against Noraristeromycin (NAM) and the three 2-substituted nucleoside inhibitors, 2-Amino-NAM, 2-Fluoro-NAM and 2-Bromo-NAM. The results showed that the inhibition of the 2-substituted nucleoside inhibitors by the mutant enzyme was significantly reduced. A similar study by Nakanishi *et al.* (2005), where *PfSAHH* mutants harboring C59T, C59S and A48Q amino acid substitutions were analyzed, also showed that activities of NAM and 2-F-NAM against the mutant enzymes were significantly reduced. Although these data describe inhibitor sensitivity to induced mutations based on the differences between the *PfSAHH* and the *H. sapiens* SAHH (*HsSAHH*), the data does not explain how and where mutations may arise on the protein, that will cause drug resistance in real life when selection pressure in the parasite is involved. Nevertheless, the absence of isoforms and paralogs as well as the identification of *PfSAHH* as a choke-point reduces the possibility of the target being resistant to drugs.

3.5.4 Toxicity

The analysis of the ortholog information on the **orthology tab** can be used to reveal whether drugs designed against *PfSAHH* can also act on similar proteins in the human host, thus causing toxicity. The data reveals that *PfSAHH* does have an ortholog in the *H. sapiens* host.

HsSAHH has an Ensembl identifier ENSP00000217426 and a UniProt accession P23526. This protein is 432 amino acids long, 47 amino acids shorter than its *P. falciparum* counterpart, which means that there are some differences between the two proteins. However, as mentioned in Section 1.3.4, orthology alone is not sufficient enough to rule out a potential target if it has an ortholog in human (Hopkins *et al.*, 2011). Selectivity and structural differences between the two SAHH enzymes must also be taken into consideration.

The differences between the *PfSAHH* and *HsSAHH* enzymes were analyzed by Bujnicki *et al.* (2003) through homology-modelling. They identified that a single amino acid substitution between Cys59 of *PfSAHH* and Thr60 of *HsSAHH* in the substrate binding site accounts for the selectivity in the binding of inhibitors between the two enzymes. However, due to the absence of an experimentally determined crystal structure of the *PfSAHH*, it was difficult to identify the structural differences in the binding sites of *PfSAHH* and *HsSAHH* that are responsible for the observed selectivity using the modelled structure of *PfSAHH*. The availability of the crystal structure of *PfSAHH* identified by Tanaka *et al.* (2004) sheds some light on how the single amino acid substitution is responsible for the selectivity of inhibitors between *PfSAHH* and *HsSAHH*.

The structural comparison of the crystal structures of the two enzymes (PDB codes 1V8B and 1LI4 for *PfSAHH* and *HsSAHH* respectively) revealed that the large -CH₃ group of Thr60 in *HsSAHH* is replaced by the small -SH group of the Cys59 in *PfSAHH*, thereby creating a space in the binding site of *PfSAHH* (Figure 3.26). The surface depression created allows for the binding of bulkier inhibitors in the binding site of *PfSAHH* that would not bind on *HsSAHH*, which lacks the surface depression due to the presence of the -CH₃ of Thr60 (Tanaka *et al.*, 2004). Although an ortholog of *PfSAHH* is present in the human host, the small differences in the substrate binding site makes it possible to design drugs that can only act on *PfSAHH* without any undesired binding to its ortholog which may cause toxicity to the human host.

3.5.5 Structural information

With structural information, *PfSAHH* is assessed for any experimentally determined crystal structures or modelled structures which can be used for *in silico* docking to ligands. *PfSAHH* does have an experimentally solved crystal structure as shown by the PDB-BLAST results

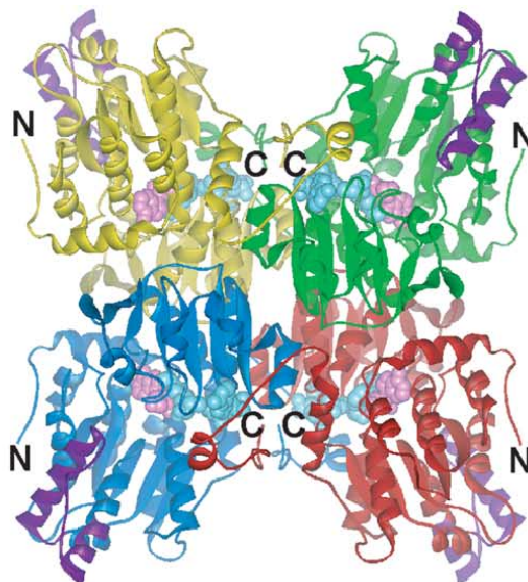


Figure 3.24: **Crystal structure of the tetrameric *PfSAHH* enzyme complexed with adenosine and NAD molecules.** The four subunits making up the tetrameric enzyme are shown in different colours (yellow, green, blue and red) with the insertion loop shown in purple. Adenosine (pink) and NAD (cyan) molecules are represented as ball-and-stick models. [Adapted from Tanaka *et al.* (2004)].

on the **structure tab**. The crystal structure of *PfSAHH* (PDB code 1V8B) was identified by molecular replacement (MR) method using the known coordinate sets from the crystal structure of the *HsSAHH* (PDB code 1LI4) (Tanaka *et al.*, 2004). The structure shows that the *PfSAHH* enzyme is a tetramer of identical subunits (Figure 3.24). Each subunit is 479 amino acids long and has a bound NAD co-factor. The subunits consist of three domains i.e., the substrate binding domain, the co-factor binding domain and the C-terminal domain (Figure 3.25).

The substrate binding domain comprises of residues 1 - 255 (258 amino acids long) and is made up of ten α -helices and six β -strands. The co-factor binding domain comprises of residues 226 - 399 (174 amino acids long) and is made up of six-stranded parallel β -sheets sandwiched between α -helices. The C-terminal domain is made up of residues 433 - 479 and is a helix-loop-helix-loop structure which extends to the adjacent subunit (Figure 3.25). The C-terminal domain covers part of the bound NAD molecule in the co-factor binding site of the adjacent subunit. Tanaka *et al.* (2004) also showed that the each subunit of the *PfSAHH* enzyme contains a 41 amino acid (146 - 168) long insert at the edge of the substrate binding site, which is not present in the mammalian SAHH.

As mentioned before, homology-modelling of *PfSAHH* and *HsSAHH* revealed that a single substitution between Cys59 of *PfSAHH* and Thr60 of *HsSAHH* accounts for inhibitor selectivity

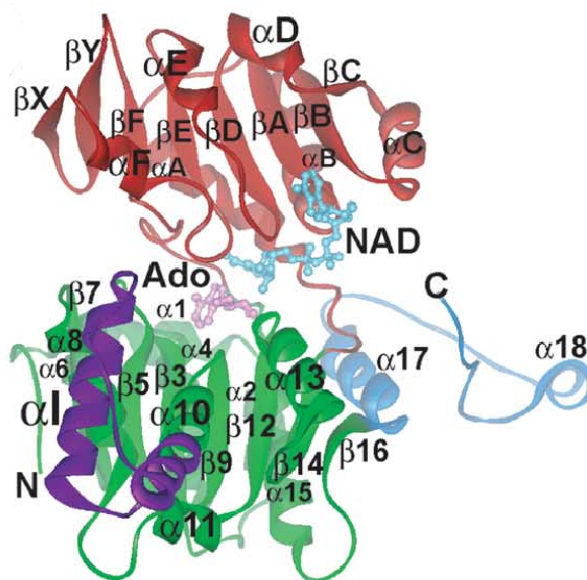


Figure 3.25: Crystal structure of the *PfSAHH* subunit complexed with adenosine and NAD. The substrate binding domain is shown in green, the co-factor binding domain is shown in red, the insertion loop is shown in purple and the C-terminal domain is colored in blue. The adenosine (pink) and NAD (cyan) molecules are represented as ball-and-stick models. [Adapted from Tanaka *et al.* (2004)].

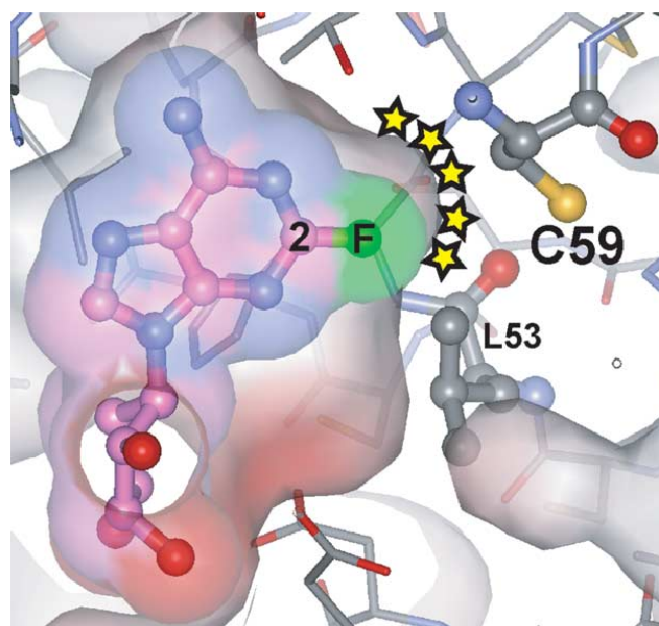


Figure 3.26: Active site of *PfSAHH*. The molecule 2-Fluoro-NAM is modelled onto the active site of *PfSAHH* to show the space (indicated by stars) created by the substitution of Thr60 for Cys59 in the active site of *PfSAHH*. [Adapted from Tanaka *et al.* (2004)].

between the two enzymes. Analysis of the active site by Tanaka *et al.* (2004) show that the -SH group of the Cys59 in *PfSAHH* occupies the large -CH₃ group position of Thr60 in *HsSAHH*, creating a space in the binding site of *PfSAHH* (Figure 3.26). The presence of the crystal

structure as well as the predicted models of the *PfSAHH* enzyme may facilitate the *in silico* rational design of drugs in cases where *in vitro* and *in vivo* methods of drugs design fail. Furthermore, the differences in the active site caused by the single substitution may be used to develop drugs that selectively inhibit *PfSAHH*.

3.5.6 Druggability

According to the results from the BLAST search against DrugEBility, *PfSAHH* was found not to be druggable. The BLAST search matched eight domains belonging to the crystal structure 1V8B of *PfSAHH* as well as its protein sequence, P50250. Other significant matches to *PfSAHH* were also not druggable. Analyzing the druggability calculations for *PfSAHH* in the DrugEBility database, we find that 2 domains, the substrate binding domain and the NAD binding domains, were used to calculate the druggability for the protein (Figure 3.28). These two domains were represented by 8 different structures (4 structures for each domain), all of which belong to the crystal structure *PfSAHH*.

Looking at the druggability calculations for the NAD binding domain, none of the sites identified were druggable. One of the sites identified occurred at the binding site of NAD. However, even though the site is known to bind NAD, it was found to be undruggable (Figure 3.27a).

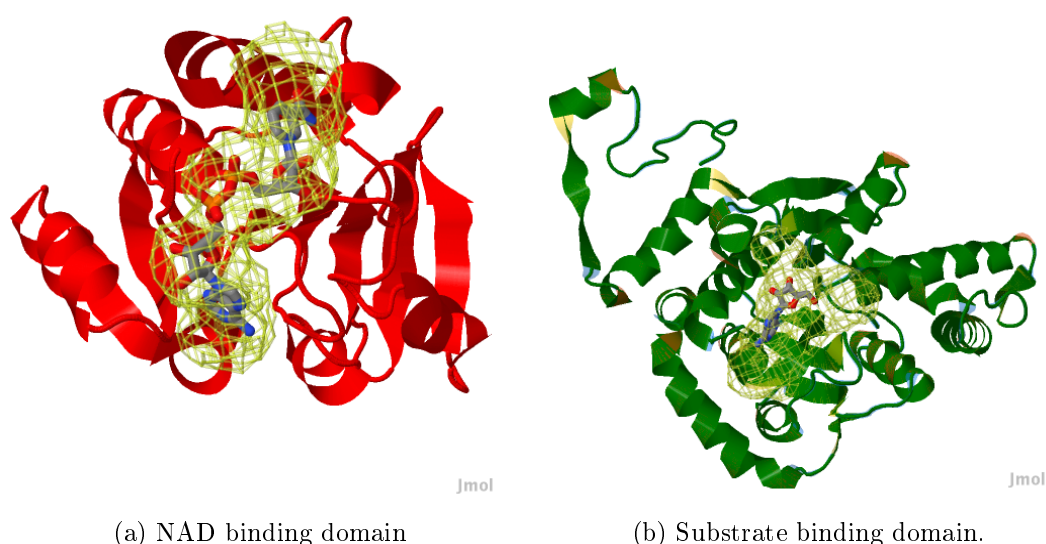



Figure 3.27: Undruggable sites identified at the known binding sites on the two *PfSAHH* domains. (a) NAD binding domain (red) complexed with the NAD molecule (CPK). (b) Substrate binding domain (green) complexed with adenosine (CPK). The predicted undruggable sites are displayed in yellow, and occur where the known molecules bind (Images produce with the Jmol applet available in DrugEBility).


EMBL-EBI [Terms of Use](#) [Privacy](#) [Cookies](#)

[Databases](#) [Tools](#) [Research](#) [Training](#) [Industry](#) [About Us](#) [Help](#) [Site Index](#)

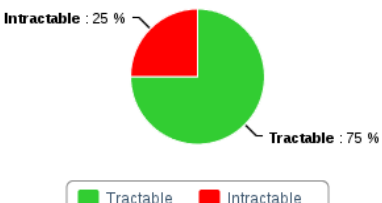
ChEMBL 

[DrugEBllity](#) **[Structure Based](#)** [Feature Based](#) [Ligand Based](#)

Overview of P50250 Adenosylhomocysteinase

Accession	P50250		
Description	Adenosylhomocysteinase		
Organism	Plasmodium falciparum		
ChEMBL ID	CHEMBL6076	Approved Drug Target	No
Pfam Domains			
Amino Acid Sequence	>P50250 MVENKSKVKDISLAPFGKMQEISENEMPLMRIREEYGDQP LKNAKITGC LHMTVECALLIETLQKLG QIRWCSCNIYSTADYAAAAVSTLEINVTVFAWKNETLEEYWCVESALTWGDGDDNGPDHIVDDGGDALLV HKGVEYEKLYEEKNILPDPEKAKNEEERCF LTLKNSILKNPKWNTIAKKIIGVSEETTTGVLR LKKMDK QNELLF TAINVNDAVTKQKYDNVYGCRRHSLPDGLMRATDF LISGKIVVICGYGDVGGCASSMKGLGARVY ITEIDPICAIQAVNEGFNVVTLDEIVDKGDFFITCTGNVDVIKLEHL LKMKNNAVVGNIGHFDEIQVNEL FNYKGIHIENVKPQVDRITLPNGNKIIVLARGRL LNLGCATGHPAFVMSFSCNQTFQAQLDLWQNKDTNKY ENKVYLLPKHLDEKVALYHLKLNASLTELDDNQCF LGVNSGPFKSNEYRY		
Structural Summaries	There are 2 structurally determined domains in this protein represented by 8 distinct structures <ul style="list-style-type: none"> 6 structures have druggable cavities via Tractability Score 0 structures have druggable cavities via Druggability Score 0 structures have druggable cavities via Ensemble Score (>0.70) [See Domain Druggability Details]		

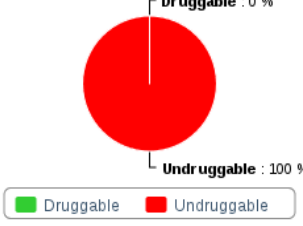
Tractability



Intractable : 25 % Tractable : 75 %

■ Tractable ■ Intractable


Druggability



Druggable : 0 % Undruggable : 100 %

■ Druggable ■ Undruggable

Ensemble Druggability



Druggable : 0 % Undruggable : 100 %

■ Druggable ■ Undruggable

[Terms of Use](#) [Privacy](#) [Cookies](#) [EBI Funding](#) [Contact EBI](#) [© European Bioinformatics Institute 2012. EBI is an Outstation of the European Molecular Biology Laboratory.](#)

Figure 3.28: Summary of the druggability calculations for *PfSAHH* (<https://www.ebi.ac.uk/chembl/drugability/protein/P50250>).

The same results are seen with the substrate binding site on *PfSAHH*; in the four sites identified, none were druggable and one occurred at the substrate binding (Figure 3.27b). However, even though the substrate binding of *PfSAHH* site was predicted not to be able to bind small molecules, there has been reports of inhibitors designed against to inhibit the enzyme, which proved to be active when tested *in vitro* and *in vivo* (Shuto *et al.*, 2002). The reasons for these confusing results could be attributed to the method used for calculating druggability of the protein.

Many factors affect the binding of molecules to a protein. The flexibility of the binding site, the overall conformation of the protein and post translational modifications are some of the factors that contribute to binding of molecules to proteins (Hopkins *et al.*, 2011). As mentioned in Section 3.5.5, *PfSAHH* is a tetrameric enzyme made up of four identical subunits. By using only domains (which are parts of the protein) in calculating druggability, the method may miss other other residues from other domains that contribute to the binding site. Other binding sites that occur in the overall surface of the protein and between domains may also be missed if only domains are used. It is thus worth noticing that binding sites do not only occur in individual domains, but rather result from the contribution of different domains and other residues in the overall conformed protein. Calculating druggability using a complete structure of the active enzyme may also be useful in order to determine the binding sites that results from the contribution of the domains and residues of the overall protein structure.

3.6 Prioritization of potential drug targets in malaria using Discovery 2.0

Analysis of *PfSAHH* as a potential target for malaria using the data in Discovery reveals that using the assessment criteria defined in this study, the protein can be used as a potential target for malaria (Table 3.8). By asking the questions like those seen in Table 3.8, guided decisions on each of the target selection criteria may be reached. These decisions determine the overall possibility of using a protein as a potential target for malaria. For *PfSAHH*, the essentiality data reveals that the protein acts on the important pathway for the parasite. Inhibition of this methionine metabolism pathway through the inhibition of the enzyme causes other methionine-

dependent reactions to stop, thereby causing death to the parasite as essential reactions also inhibited. This information is supported by literature as well as the identification of *PfSAHH* as a choke-point.

The presence of expression, purification and activity assays for the enzyme is advantageous to the drug discovery process since the protein can be readily obtained in large amounts and activity of the enzyme can be easily detected when different small molecules are tested against it in HTS techniques. Assessing the potential for resistance and toxicity with the data in Discovery 2.0 also yields positive results. No paralogs or isoforms for *PfSAHH* were identified within *P. falciparum*, and since the enzyme is a choke-point, potential of resistance is less possible. Even though the protein has an ortholog in human, structural analysis between the two SAHH enzymes reveal that there are significant differences which account for selectivity of inhibitors. However, even though the druggability data may have negative results, the protein does have small molecules that have been found to inhibit the enzyme. The summary for the assessment of *PfSAHH* as a drug target for malaria is shown in Table 3.8.

The downfall of the method of assessing protein targets described here, however, is that the

Table 3.8: Summary for the assessment of *PfSAHH* as a drug target.

	Assessment criteria	Yes(✓)/No(X)
Essentiality	Is protein vital for survival?	✓
	Choke-point?	✓
	Literature evidence?	✓
Assay feasibility	Can be readily expressed?	✓
	Activity assay present?	✓
	Literature evidence?	✓
Resistance	Isoforms present?	X
	Paralogs present?	X
	Choke-point?	✓
	Literature evidence?	✓
Toxicity	Ortholog in human present?	✓
	Differences in binding sites?	✓
	Literature evidence?	✓
Structural info.	Experimental structure present?	✓
	Predicted structure present?	✓
	Known binding sites (literature)?	✓
Druggability	Is protein druggable?	X
	Any druggable domain?	X
	Method of calculation (whole protein?)	X
	Small molecules/inhibitors reported (literature)?	✓

user does not have an immediate decision as to whether the protein can be used as a target or not, and literature may have to be consulted before a decision can be made. For future improvements on Discovery 2.0, a “target scoring function” could be incorporated, which will score a protein based on its properties and data available in the resource. The questions that were used in this study to assess a protein as a target may be used to construct a decision making pipeline that will determine whether a protein may be used as a drug target. The pipeline will make decisions based on the relevance of each target assessment criteria and protein properties that are more important than the others when it comes to prioritization of drug targets. An overall score will be derived to determine if a protein can be used as a drug target. Other data, such as choke-point analysis from literature, could also be incorporated in future improvements of the resource to aid in decision making of drug target prioritization.

3.7 Discussion

Discovery 2.0 can be accessed at <http://discovery.bi.up.ac.za/>, where a protein identifier (Ensembl or PlasmoDB), UniProt accession or a protein name can be used to search for a protein. A more advanced search is also available, where users can filter out protein sequences in a step-by-step manner using different filters available. With the **advanced search**, users may set the parameters of each filter and choose to combine filters using the “AND” option or exclude some results using the “AND NOT” options. The results for each protein are returned in a tabbed environment, with each tab displaying a different category of data which can be used to get information about the protein, and also to assess a protein as a drug target.

Discovery 2.0 currently contains 140 218 protein sequences belonging to *A. gambiae*, *H. sapiens*, *P. berghei*, *P. chabaudi*, *P. falciparum*, *P. knowlesi*, *P. vivax* and *P. yoelii*. Assignment of the protein sequences with UniProt accessions made it possible to download, store and map data from different databases onto proteins in the Discovery 2.0 database. The mapping of GO terms, metabolic pathways, interactions and druggability data to proteins was possible through the UniProt accessions. However, some annotation data in Discovery 2.0 is not downloaded and stored in the database, but rather obtained as the user accesses the information on a particular protein as seen with metabolic pathways and EC numbers assignment (Section 3.3 and Figure 3.12).

The case studies presented in this chapter demonstrate how users can predefine a set of criteria to be met by a protein and how to use these criteria to filter out proteins in a step-by-step manner using the **advanced search** functionality of Discovery 2.0, in order to find the desired protein(s). Users can also split their searches into two or more branches in order to apply different filters on their search results. Basic information about the protein can be obtained from the tabs representing different categories of annotation data. These tabs are “**Summary**”, “**Function**”, “**Gene ontology**”, “**Orthology**”, “**Structure**”, “**Metabolic pathways**”, “**Protein-ligand interactions**”, “**Druggability**”, “**Expression**” and “**Literature**”. The different categories of annotation data do not only give basic information about the protein, but can also be used to assess a protein as a potential drug target for malaria.

Assessing the protein *PfSAHH* as a potential drug target demonstrated that Discovery 2.0 resource is a useful tool for prioritizing drug targets when used in combination with literature. However, this requires a lot of work and one cannot make immediate decisions on whether a protein can be a good malaria drug target or not. Incorporating a statistical model into Discovery 2.0 that will score a protein as a potential drug target based on the annotation data available is one of the future improvements plan for the resource.

Chapter 4

Concluding discussion

The Discovery 2.0 resource has now been populated with new features and contains new datasets that were not available in the old version of Discovery. Switching of the programming language from Python to Java has led to the resource being faster, in terms of searches, as well as to allow the incorporation of Java applets that are useful in the analysis of data. The resource is also updated automatically and provides links to other resources/databases where data was obtained. Users can query the database for a protein sequence through the **basic search** using a protein identifier, UniProt accession or a protein name. A built-in auto-complete feature of the **basic search** assist users as they enter the characters in the search space for protein identifiers and UniProt accessions. The new **advanced search** function in Discovery 2.0 allows for fast filtering of proteins based on different protein annotations. The filters available for filtering protein sequences according to annotations are *function*, *gene ontology*, *MODBASE structures*, *organism*, *orthology*, *PDB-BLAST*, *protein-ligand interactions*, *protein name* and *related PubMed articles* filters. The advantage of the **advanced search** is that it allows users to identify unknown proteins based on the properties the user want met.

The protein annotation data available in the current version of Discovery 2.0 was motivated by the different target selection criteria mentioned in this study. Currently, Discovery 2.0 contains annotation data for 140 218 protein sequences belonging to *A. gambiae*, *H. sapiens*, *P. berghei*, *P. chabaudi*, *P. falciparum*, *P. knowlesi*, *P. vivax* and *P. yoelii*. The datasets available have been extended to include UniProt accessions, links to databases where data was obtained, pathways from the MPMP and Reactome databases, GO data from UniProt-GOA, protein-protein interaction data from IntAct as well as druggability data from the DrugEBility

resource hosted by the ChEMBL database at the EBI. The data for each protein is organized in a tabbed environment (“**Summary**”, “**Function**”, “**Gene ontology**”, “**Orthology**”, “**Structure**”, “**Metabolic pathways**”, “**Protein-ligand interactions**”, “**Druggability**”, “**Expression**” and “**Literature**”) within the user-friendly web interface of Discovery 2.0. Java applets are integrated into the system for visual analysis of alignments (Jalview) and crystal structures (Jmol).

The case studies presented in this study demonstrate how users can use the **advanced search** functionality of Discovery 2.0 to identify desired proteins. A set of criteria to be met by a protein that a user is interested in can be defined, and these criteria can be applied on the filters of the **advanced search** to identify the protein. The more basic filters (*MODBASE structures*, *organism*, *protein name* and *related PubMed articles filters*) are quite simple to use and do not require advanced settings when used. However, the more advanced filters (*function*, *gene ontology*, *orthology* and *PDB-BLAST filters*) require some knowledge beforehand for the users to apply to their searches. The *function filter* match a user input to the descriptions of the InterProScan results. These are short descriptions of the contributing signatures of an InterPro entry and not the full name of the predicted family, domain or functional site. It is thus useful for a user to visit external databases (InterPro in this case) in order to identify the short description of signatures that contribute to the family, domain or functional site that they want to include in their advanced search. The *gene ontology filter* match a user input to the descriptions of the GO terms in the database. It is also useful for a user to visit the Gene Ontology database in order to identify the GO term description to use in this filter. The *orthology filter* requires a user to specify the organisms that a protein should have an ortholog of as well as the %change cut-off value. The %change is a measure of how different the protein sequence is to the query protein. The *PDB-BLAST filter* requires an *E*-value cut-off to be specified.

Looking at the results from the **advanced search** case studies, we can see how applying the filters reduces the number of protein sequences from the start to end of the step-by-step search. Each filter reduces the number of protein sequences returned. As the filters become more complex, more protein sequences are excluded and the specific proteins sequences matching the filter remain. The proteins identified in these case studies are well annotated, even though

most lack the annotation for protein-protein interactions. The links to the external databases provided for the proteins identified provide more detailed information that is not available in Discovery 2.0. The case studies do not only demonstrate how users can use the resource to search for proteins, but it also demonstrates how basic information on a protein can be obtained by analyzing the annotation data available in Discovery 2.0. However, when it comes to deciding whether a protein can be used as a potential drug target in malaria, more advanced analysis of the data, in combination with literature, is required as demonstrated by the analysis of the enzyme *PfSAHH*.

Using the data available in Discovery 2.0 together with literature for validation, the *PfSAHH* enzyme was assessed for essentiality, assay feasibility, resistance, toxicity, structural information and druggability in order to identify whether it can be used as a potential drug target for malaria. Essentiality data reveals that *PfSAHH* is involved in the metabolism of methionine, where it hydrolyses SAH to adenosine and L-homocysteine. No other enzyme is responsible for this reaction. Inhibition of *PfSAHH* causes accumulation of cellular SAH which eventually leads to cell death. This makes *PfSAHH* essential to the survival of the parasite as it is a key regulator for reactions involving methionine. The assay feasibility data reveals that the *PfSAHH* enzyme can be readily expressed in *E. coli* cells and activity assays have been described. This can greatly facilitate the drug discovery process since the protein can be obtained in large, pure quantities and activity assays for HTS are available.

Analyzing the data for possible resistance that might arise if *PfSAHH* is used as a drug target for malaria, we find that no isoform or paralogs are found in the parasite that might replace the role of *PfSAHH* if inhibited. We also find that *PfSAHH* is involved in a choke-point reaction. Thus, *PfSAHH* would make a good drug target since there is less possibility of drug resistance that might arise. However, the drug resistance that might arise through mutations cause by selection pressure also needs to be taken into account. The toxicity data reveals that *PfSAHH* has an ortholog in human, and this may cause drugs designed to act on *PfSAHH* to also act on the human ortholog, thereby causing toxicity in human. However, there are significant differences in the binding sites of these orthologs which makes it possible to design drugs that can only act on *PfSAHH* without any undesired binding to the human ortholog, thus reducing the possibility of toxicity to humans.

The structural information for *PfSAHH* shows that the enzyme has an experimentally determined crystal structure in addition to the two modelled structures. The presence of the crystal structure for this enzyme makes it possible to carry out *in silico* dockings of small molecules as well as the design/optimization of drugs when *in vitro* or *in vivo* methods are not possible. According to the druggability data, *PfSAHH* was found not to be druggable. However, small molecules that are active against the enzyme have been reported. The druggability calculations were done on the domains of the enzyme but not the whole protein. It is thus also important to calculate druggability using the structure of the whole enzyme as some binding sites that result from the contribution of all domains and other residues of the protein might have been missed when using only the domains.

The data collected from Discovery 2.0 for *PfSAHH* on essentiality, assay feasibility, resistance, toxicity, structural information and druggability does not only reveal that *PfSAHH* can be used as a drug target for malaria, but also demonstrates how Discovery 2.0 can be used to assess and prioritize proteins as drug targets. The disadvantage of the method of assessing a protein as a drug target in Discovery 2.0, however, is that the user is not given information as to whether the protein can be used as a drug target for malaria or not. Another disadvantage is that the user has to consult literature before coming to a conclusion of whether the protein they are interested in has a potential of being a drug target. This can be problematic especially when it comes to proteins that have not yet been studied. Future improvements on Discovery 2.0 include incorporating a “target scoring function” to assess proteins as drug targets based on the annotation data available. The target scoring function can be built on a decision making pipeline, where the available protein annotations under each target assessment criteria are scored based on the relevance and importance when it comes to target selection. Other improvements include incorporating choke-point analysis data into the system, which could aid in the assessment of drug targets.

Discovery 2.0 has advantages over other resources that are relevant to the drug target selection in malaria, for example, the TDR Targets database. Both the TDR Targets database and Discovery 2.0 contain information on orthology, functional annotation, metabolic pathways, structure, literature, expression data, and druggability. However, the protein-protein interaction data and UniProt accessions are not included in the TDR Targets database. Discovery 2.0

does not only provide protein information for the the *Plasmodium* species, but it also provides the human and mosquito protein information. This allows researches to do analysis in a species-comparative environment. In addition to the host/vector information provided in Discovery 2.0, the advanced search functionality allows user to do advanced queries on the database.

Using the data in the Discovery 2.0 resource, potential malaria drug targets could be identified and validated as seen with the *PfSAHH* enzyme used as an example in this study, as well as the case studies carried out using the **advanced search** functionality. Having such a system for prioritizing drug targets in a species comparative manner could reduce the time taken to identify drug targets as well as reduce the failure rates of the drug discovery process. Computational biology and bioinformatics techniques are very useful tools for translating the amount of data available from sequencing the genomes of many organisms into meaningful data that can be used to answer many biological questions that would sometimes take years to answer using traditional biological techniques. Proper utilization and mining of this data, together with experimental data from published literature could increase our knowledge on how biological systems function. Knowing how biological systems function, especially disease causing parasites, could help us identify the most important areas where attention should be focused in order to design drugs that are active against diseases affecting the human population.

Bibliography

- Adjuik, M., Babiker, A., Garner, P., Olliaro, P., Taylor, W., White, N. and Group, I. A. S. (2004) Artesunate combinations for treatment of malaria: meta-analysis. *Lancet* **363**, 9402, 9–17.
- Afonso, A., Hunt, P., Cheesman, S., Alves, A. C., Cunha, C. V., do Rosário, V. and Cravo, P. (2006) Malaria parasites can develop stable resistance to artemisinin but lack mutations in candidate genes *atp6* (encoding the sarcoplasmic and endoplasmic reticulum Ca²⁺ ATPase), *tctp*, *mdr1*, and *cg10*. *Antimicrob Agents Chemother* **50**, 2, 480–489.
- Agüero, F., Al-Lazikani, B., Aslett, M., Berriman, M., Buckner, F. S., Campbell, R. K., Carmona, S., Carruthers, I. M., Chan, A. W. E., Chen, F., Crowther, G. J., Doyle, M. A., Hertz-Fowler, C., Hopkins, A. L., McAllister, G., Nwaka, S., Overington, J. P., Pain, A., Paolini, G. V., Pieper, U., Ralph, S. A., Riechers, A., Roos, D. S., Sali, A., Shanmugam, D., Suzuki, T., Voorhis, W. C. V. and Verlinde, C. L. M. J. (2008) Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat Rev Drug Discov* **7**, 11, 900–907.
- Al-Lazikani, B., Gaulton, A., Paolini, G., Lanfear, J., Overington, J. and Hopkins, A. (2008) *The Molecular Basis of Predicting Druggability* 1315–1334 Wiley-VCH Verlag GmbH ISBN 9783527619368.
- Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J. and Zdobnov, E. M. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* **29**, 1, 37–40.

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 1, 25–29.
- Atkinson, H. J., Babbitt, P. C. and Sajid, M. (2009) The global cysteine peptidase landscape in parasites. *Trends Parasitol* **25**, 12, 573–581.
- Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Miller, J. A., Nayak, V., Pennington, C., Pinney, D. F., Roos, D. S., Ross, C., Stoeckert, C. J., Treatman, C. and Wang, H. (2009) PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* **37**, Database issue, D539–D543.
- Bahl, A., Brunk, B., Coppel, R. L., Crabtree, J., Diskin, S. J., Fraunholz, M. J., Grant, G. R., Gupta, D., Huestis, R. L., Kissinger, J. C., Labo, P., Li, L., McWeeney, S. K., Milgram, A. J., Roos, D. S., Schug, J. and Stoeckert, C. J. (2002) PlasmoDB: the *Plasmodium* genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished). *Nucleic Acids Res* **30**, 1, 87–90.
- Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res* **28**, 1, 304–305.
- Baldwin, J., Michnoff, C. H., Malmquist, N. A., White, J., Roth, M. G., Rathod, P. K. and Phillips, M. A. (2005) High-throughput screening for potent and selective inhibitors of *Plasmodium falciparum* dihydroorotate dehydrogenase. *J Biol Chem* **280**, 23, 21847–21853.
- Baniecki, M. L., Wirth, D. F. and Clardy, J. (2007) High-throughput *Plasmodium falciparum* growth assay for malaria drug discovery. *Antimicrob Agents Chemother* **51**, 2, 716–723.
- Barthelme, J., Ebeling, C., Chang, A., Schomburg, I. and Schomburg, D. (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res* **35**, Database issue, D511–D514.
- Belmont, M., Cazzamali, G., Williamson, M., Hauser, F. and Grimmelikhuijzen, C. J. P. (2006)

- Identification of four evolutionarily related G protein-coupled receptors from the malaria mosquito *Anopheles gambiae*. *Biochem Biophys Res Commun* **344**, 1, 160–165.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 1, 235–242.
- Blandin, S., Moita, L. F., Köcher, T., Wilm, M., Kafatos, F. C. and Levashina, E. A. (2002) Reverse genetics in the mosquito *Anopheles gambiae*: targeted disruption of the Defensin gene. *EMBO Rep* **3**, 9, 852–856.
- Bockaert, J. and Pin, J. P. (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J* **18**, 7, 1723–1729.
- Bornman, R., de Jager, C., Worku, Z., Farias, P. and Reif, S. (2010) DDT and urogenital malformations in newborn boys in a malarial area. *BJU Int* **106**, 3, 405–411.
- Bouwman, H., van den Berg, H. and Kylin, H. (2011) DDT and malaria prevention: addressing the paradox. *Environ Health Perspect* **119**, 6, 744–747.
- Bréhélin, L., Dufayard, J.-F. and Gascuel, O. (2008) PlasmoDraft: a database of *Plasmodium falciparum* gene function predictions based on postgenomic data. *BMC Bioinformatics* **9**, 440.
- Bruce-Chwatt, L. J. (1981) Alphonse Laveran's discovery 100 years ago and today's global fight against malaria *Journal of the Royal Society of Medicine* **74**, 531 – 536.
- Buchholz, K., Burke, T. A., Williamson, K. C., Wiegand, R. C., Wirth, D. F. and Marti, M. (2011) A high-throughput screen targeting malaria transmission stages opens new avenues for drug development. *J Infect Dis* **203**, 10, 1445–1453.
- Bujnicki, J. M., Prigge, S. T., Caridha, D. and Chiang, P. K. (2003) Structure, evolution, and inhibitor interaction of S-adenosyl-L-homocysteine hydrolase from *Plasmodium falciparum*. *Proteins* **52**, 4, 624–632.
- Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Hub, A. O. and Group, W. P. W. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 2, 288–289.

- Ceol, A., Aryamontri, A. C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L. and Cesareni, G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* **38**, Database issue, D532–D539.
- Chang, A., Scheer, M., Grote, A., Schomburg, I. and Schomburg, D. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res* **37**, Database issue, D588–D592.
- Chaudhuri, R., Ahmed, S., Ansari, F. A., Singh, H. V. and Ramachandran, S. (2008) MalVac: database of malarial vaccine candidates. *Malar J* **7**, 184.
- Chen, F., Mackey, A. J., Stoeckert, C. J. and Roos, D. S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**, Database issue, D363–D368.
- Chen, Y.-P. P. and Chen, F. (2008) Identifying targets for drug discovery using bioinformatics. *Expert Opin Ther Targets* **12**, 4, 383–389.
- Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., Salzberg, A. C. and Huang, E. S. (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* **25**, 1, 71–75.
- Coleman, R. G., Salzberg, A. C. and Cheng, A. C. (2006) Structure-based identification of small molecule binding sites using a free energy model. *J Chem Inf Model* **46**, 6, 2631–2637.
- Consortium, U. (2008) The universal protein resource (UniProt). *Nucleic Acids Res* **36**, Database issue, D190–D195.
- Consortium, U. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**, Database issue, D71–D75.
- Coombs, G. H., Goldberg, D. E., Klemba, M., Berry, C., Kay, J. and Mottram, J. C. (2001) Aspartic proteases of *Plasmodium falciparum* and other parasitic protozoa as drug targets. *Trends Parasitol* **17**, 11, 532–537.
- Coombs, G. H. and Mottram, J. C. (1997) Parasite proteinases and amino acid metabolism: possibilities for chemotherapeutic exploitation. *Parasitology* **114** Suppl, S61–S80.

- Copeland, R. A., Davis, J. P., Dowling, R. L., Lombardo, D., Murphy, K. B. and Patterson, T. A. (1995) Recombinant human dihydroorotate dehydrogenase: expression, purification, and characterization of a catalytically functional truncated enzyme. *Arch Biochem Biophys* **323**, 1, 79–86.
- Cowman, A. F., Galatis, D. and Thompson, J. K. (1994) Selection for mefloquine resistance in *Plasmodium falciparum* is linked to amplification of the *pfmdr1* gene and cross-resistance to halofantrine and quinine. *Proc Natl Acad Sci U S A* **91**, 3, 1143–1147.
- Creedon, K. A., Rathod, P. K. and Wellem, T. E. (1994) *Plasmodium falciparum* S-adenosylhomocysteine hydrolase. cDNA identification, predicted protein sequence, and expression in *Escherichia coli*. *J Biol Chem* **269**, 23, 16364–16370.
- Curran, S. P. and Ruvkun, G. (2007) Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet* **3**, 4, e56.
- Dash, C., Kulkarni, A., Dunn, B. and Rao, M. (2003) Aspartic peptidase inhibitors: implications in drug development. *Crit Rev Biochem Mol Biol* **38**, 2, 89–119.
- de Angelis, M. H. H., Flaswinkel, H., Fuchs, H., Rathkolb, B., Soewarto, D., Marschall, S., Heffner, S., Pargent, W., Wuensch, K., Jung, M., Reis, A., Richter, T., Alessandrini, F., Jakob, T., Fuchs, E., Kolb, H., Kremmer, E., Schaeble, K., Rollinski, B., Roscher, A., Peters, C., Meitinger, T., Strom, T., Steckler, T., Holsboer, F., Klopstock, T., Gekeler, F., Schindewolf, C., Jung, T., Avraham, K., Behrendt, H., Ring, J., Zimmer, A., Schughart, K., Pfeffer, K., Wolf, E. and Balling, R. (2000) Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat Genet* **25**, 4, 444–447.
- Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., Bely, B., Browne, P., Chan, W. M., Eberhardt, R., Gardner, M., Laiho, K., Legge, D., Magrane, M., Pichler, K., Poggioli, D., Sehra, H., Auchincloss, A., Axelsen, K., Blatter, M.-C., Boutet, E., Braconi-Quintaje, S., Breuza, L., Bridge, A., Coudert, E., Estreicher, A., Famiglietti, L., Ferro-Rojas, S., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., James, J., Jimenez, S., Jungo, F., Keller, G., Lemercier, P., Lieberherr, D., Masson, P., Moinat, M., Pedruzzi, I., Poux, S., Rivoire, C., Roechert, B., Schneider, M., Stutz, A.,

- Sundaram, S., Tognolli, M., Bougueleret, L., Argoud-Puy, G., Cusin, I., Duek-Roggli, P., Xenarios, I. and Apweiler, R. (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* **40**, Database issue, D565–D570.
- Djimdé, A., Doumbo, O. K., Cortese, J. F., Kayentao, K., Doumbo, S., Diourté, Y., Dicko, A., Su, X. Z., Nomura, T., Fidock, D. A., Wellems, T. E. and Plowe, C. V. (2001) A molecular marker for chloroquine-resistant *falciparum* malaria. *N Engl J Med* **344**, 4, 257–263.
- Dondorp, A. M., Nosten, F., Yi, P., Das, D., Phyto, A. P., Tarning, J., Lwin, K. M., Ariey, F., Hanpithakpong, W., Lee, S. J., Ringwald, P., Silamut, K., Imwong, M., Chotivanich, K., Lim, P., Herdman, T., An, S. S., Yeung, S., Singhasivanon, P., Day, N. P. J., Lindergardh, N., Socheat, D. and White, N. J. (2009) Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med* **361**, 5, 455–467.
- Doyle, M. A., Gasser, R. B., Woodcroft, B. J., Hall, R. S. and Ralph, S. A. (2010) Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* **11**, 222.
- Evans, M. J., Saghatelian, A., Sorensen, E. J. and Cravatt, B. F. (2005) Target discovery in small-molecule cell-based screens by in situ proteome reactivity profiling. *Nat Biotechnol* **23**, 10, 1303–1307.
- Fatumo, S., Plaimas, K., Mallm, J.-P., Schramm, G., Adebisi, E., Oswald, M., Eils, R. and König, R. (2009) Estimating novel potential drug targets of *Plasmodium falciparum* by analysing the metabolic network of knock-out strains *in silico*. *Infect Genet Evol* **9**, 3, 351–358.
- Fidock, D. A., Nomura, T., Talley, A. K., Cooper, R. A., Dzekunov, S. M., Ferdig, M. T., Ursos, L. M., Sidhu, A. B., Naudé, B., Deitsch, K. W., Su, X. Z., Wootton, J. C., Roepe, P. D. and Wellems, T. E. (2000) Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol Cell* **6**, 4, 861–871.
- Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., Shanmugam, D., Roos, D. S. and Stoeckert, C. J. (2011) Using OrthoMCL to assign proteins to OrthoMCL-

DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6.12.1–Unit 6.1219.

Fogel, G. B., Cheung, M., Pittman, E. and Hecht, D. (2008) *In silico* screening against wild-type and mutant *Plasmodium falciparum* dihydrofolate reductase. *J Mol Graph Model* **26**, 7, 1145–1152.

Fredriksson, R. and Schiöth, H. B. (2005) The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* **67**, 5, 1414–1425.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M. A., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. and Barrell, B. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 6906, 498–511.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. and Bairoch, A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 13, 3784–3788.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. and Overington, J. P. (2011) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*.

Ginsburg, H. (2006) Progress in *in silico* functional genomics: the malaria Metabolic Pathways database. *Trends Parasitol* **22**, 6, 238–240.

Ginsburg, H. (2009) Caveat emptor: limitations of the automated reconstruction of metabolic pathways in *Plasmodium*. *Trends in Parasitology* **25**, 1, 37 – 43 ISSN 1471-4922.

Goldberg, D. E. (2002) Parasitology. When the host is smarter than the parasite. *Science* **296**, 5567, 482–483.

- Hajduk, P. J., Huth, J. R. and Fesik, S. W. (2005) Druggability indices for protein targets derived from NMR-based screening data. *J Med Chem* **48**, 7, 2518–2525.
- Haw, R., Hermjakob, H., D'Eustachio, P. and Stein, L. (2011) Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics* **11**, 18, 3598–3613.
- Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M. C., Wides, R., Salzberg, S. L., Loftus, B., Yandell, M., Majoros, W. H., Rusch, D. B., Lai, Z., Kraft, C. L., Abril, J. F., Anthouard, V., Arensburger, P., Atkinson, P. W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chaturverdi, K., Christophides, G. K., Chrystal, M. A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C. A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M. E., Hladun, S. L., Hogan, J. R., Hong, Y. S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J.-J., Lobo, N. F., Lopez, J. R., Malek, J. A., McIntosh, T. C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S. D., O'Brochta, D. A., Pfannkoch, C., Qi, R., Regier, M. A., Remington, K., Shao, H., Sharakhova, M. V., Sitter, C. D., Shetty, J., Smith, T. J., Strong, R., Sun, J., Thomasova, D., Ton, L. Q., Topalis, P., Tu, Z., Unger, M. F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K. J., Wortman, J. R., Wu, M., Yao, A., Zdobnov, E. M., Zhang, H., Zhao, Q., Zhao, S., Zhu, S. C., Zhimulev, I., Coluzzi, M., della Torre, A., Roth, C. W., Louis, C., Kalush, F., Mural, R. J., Myers, E. W., Adams, M. D., Smith, H. O., Broder, S., Gardner, M. J., Fraser, C. M., Birney, E., Bork, P., Brey, P. T., Venter, J. C., Weissenbach, J., Kafatos, F. C., Collins, F. H. and Hoffman, S. L. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 5591, 129–149.
- Hopkins, A. L., Bickerton, G. R., Carruthers, I. M., Boyer, S. K., Rubin, H. and Overington, J. (2011) Rapid analysis of pharmacology for infectious diseases. *Curr Top Med Chem* **11**, 10, 1292–1300.
- Hopkins, A. L. and Groom, C. R. (2002) The druggable genome. *Nat Rev Drug Discov* **1**, 9, 727–730.

- Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. and Flicek, P. (2009) Ensembl 2009. *Nucleic Acids Res* **37**, Database issue, D690–D697.
- Hunt, S. Y., Detering, C., Varani, G., Jacobus, D. P., Schiehser, G. A., Shieh, H.-M., Nevchas, I., Terpinski, J. and Sibley, C. H. (2005) Identification of the optimal third generation antifolate against *P. falciparum* and *P. vivax*. *Mol Biochem Parasitol* **144**, 2, 198–205.
- Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H. and Yeats, C. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, Database issue, D211–D215.
- Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J. A., Scheremetjew, M., Tate, J., Thimmajananathan, M., Thomas, P. D., Wu, C. H., Yeats, C. and Yong, S.-Y. (2012) InterPro in 2011: New developments in the family and domain prediction database *Nucleic Acids Research* **40**, Database issue, D306–D312.

- Huthmacher, C., Hoppe, A., Bulik, S. and Holzhütter, H.-G. (2010) Antimalarial drug targets in *Plasmodium falciparum* predicted by stage-specific metabolic network analysis. *BMC Syst Biol* **4**, 120.
- Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E. and Stein, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33**, Database issue, D428–D432.
- Joubert, F., Harrison, C. M., Koegelenberg, R. J., Odendaal, C. J. and de Beer, T. A. P. (2009) Discovery: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. *Malar J* **8**, 178.
- Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, Database issue, D480–D484.
- Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 1, 27–30.
- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34**, Database issue, D354–D357.
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, Database issue, D109–D114.
- Kasam, V., Salzemann, J., Botha, M., Dacosta, A., Degliesposti, G., Isea, R., Kim, D., Maass, A., Kenyon, C., Rastelli, G., Hofmann-Apitius, M. and Breton, V. (2009) WISDOM-II: screening against multiple targets implicated in malaria using computational grid infrastructures. *Malar J* **8**, 88.
- Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert,

- B., Orchard, S. and Hermjakob, H. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**, Database issue, D841–D846.
- Kissinger, J. C., Brunk, B. P., Crabtree, J., Fraunholz, M. J., Gajria, B., Milgram, A. J., Pearson, D. S., Schug, J., Bahl, A., Diskin, S. J., Ginsburg, H., Grant, G. R., Gupta, D., Labo, P., Li, L., Mailman, M. D., McWeeney, S. K., Whetzel, P., Stoeckert, C. J. and Roos, D. S. (2002) The *Plasmodium* genome database. *Nature* **419**, 6906, 490–492.
- Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., Campbell, K. S., Christophides, G. K., Christley, S., Dialynas, E., Hammond, M., Hill, C. A., Konopinski, N., Lobo, N. F., MacCallum, R. M., Madey, G., Megy, K., Meyer, J., Redmond, S., Severson, D. W., Stinson, E. O., Topalis, P., Birney, E., Gelbart, W. M., Kafatos, F. C., Louis, C. and Collins, F. H. (2009) VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res* **37**, Database issue, D583–D587.
- Li, L., Stoeckert, C. J. and Roos, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 9, 2178–2189.
- Lindsay, M. A. (2003) Target discovery. *Nat Rev Drug Discov* **2**, 10, 831–838.
- Longo, M., Zanoncelli, S., Manera, D., Brughera, M., Colombo, P., Lansén, J., Mazué, G., Gomes, M., Taylor, W. R. J. and Olliaro, P. (2006) Effects of the antimalarial drug dihydroartemisinin (DHA) on rat embryos in vitro. *Reprod Toxicol* **21**, 1, 83–93.
- Magrane, M. and Consortium, U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009.
- Manning, G., Whyte, D. B., Martinez, R., Hunter, T. and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science* **298**, 5600, 1912–1934.
- McGowan, S., Oellig, C. A., Birru, W. A., Caradoc-Davies, T. T., Stack, C. M., Lowther, J., Skinner-Adams, T., Mucha, A., Kafarski, P., Grembecka, J., Trenholme, K. R., Buckle, A. M., Gardiner, D. L., Dalton, J. P. and Whisstock, J. C. (2010) Structure of the *Plasmodium falciparum* M17 aminopeptidase and significance for the design of drugs targeting the neutral exopeptidases. *Proc Natl Acad Sci U S A* **107**, 6, 2449–2454.

- Miranda-Saavedra, D. and Barton, G. J. (2007) Classification and functional annotation of eukaryotic protein kinases. *Proteins* **68**, 4, 893–914.
- Mittl, P. R. and Grütter, M. G. (2006) Opportunities for structure-based design of protease-directed drugs. *Curr Opin Struct Biol* **16**, 6, 769–775.
- Mushegian, A. R. and Koonin, E. V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93**, 19, 10268–10273.
- Nakanishi, M., Iwata, A., Yatome, C. and Kitade, Y. (2001) Purification and properties of recombinant *Plasmodium falciparum* S-adenosyl-L-homocysteine hydrolase. *J Biochem* **129**, 1, 101–105.
- Nakanishi, M., Yabe, S., Tanaka, N., Ito, Y., Nakamura, K. T. and Kitade, Y. (2005) Mutational analyses of *Plasmodium falciparum* and human S-adenosylhomocysteine hydrolases. *Mol Biochem Parasitol* **143**, 2, 146–151.
- Nguyen, C., Kasinathan, G., Leal-Cortijo, I., Musso-Buendia, A., Kaiser, M., Brun, R., Ruiz-Pérez, L. M., Johansson, N. G., González-Pacanowska, D. and Gilbert, I. H. (2005) Deoxyuridine triphosphate nucleotidohydrolase as a potential antiparasitic drug target. *J Med Chem* **48**, 19, 5942–5954.
- Nguyen, C., Ruda, G. F., Schipani, A., Kasinathan, G., Leal, I., Musso-Buendia, A., Kaiser, M., Brun, R., Ruiz-Pérez, L. M., Sahlberg, B.-L., Johansson, N. G., Gonzalez-Pacanowska, D. and Gilbert, I. H. (2006) Acyclic nucleoside analogues as inhibitors of *Plasmodium falciparum* dUTPase. *J Med Chem* **49**, 14, 4183–4195.
- Noedl, H., Se, Y., Schaefer, K., Smith, B. L., Socheat, D., Fukuda, M. M. and in Cambodia 1 (ARC1) Study Consortium, A. R. (2008) Evidence of artemisinin-resistant malaria in western Cambodia. *N Engl J Med* **359**, 24, 2619–2620.
- Nolan, P. M., Peters, J., Strivens, M., Rogers, D., Hagan, J., Spurr, N., Gray, I. C., Vitor, L., Brooker, D., Whitehill, E., Washbourne, R., Hough, T., Greenaway, S., Hewitt, M., Liu, X., McCormack, S., Pickford, K., Selley, R., Wells, C., Tymowska-Lalanne, Z., Roby, P., Glenister, P., Thornton, C., Thaung, C., Stevenson, J. A., Arkell, R., Mburu, P., Hardisty, R.,

- Kiernan, A., Erven, A., Steel, K. P., Voegelings, S., Guenet, J. L., Nickols, C., Sadri, R., Nasse, M., Isaacs, A., Davies, K., Browne, M., Fisher, E. M., Martin, J., Rastan, S., Brown, S. D. and Hunter, J. (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat Genet* **25**, 4, 440–443.
- Nosten, F., van Vugt, M., Price, R., Luxemburger, C., Thway, K. L., Brockman, A., McGready, R., ter Kuile, F., Looareesuwan, S. and White, N. J. (2000) Effects of artesunate-mefloquine combination on incidence of *Plasmodium falciparum* malaria and mefloquine resistance in western Thailand: a prospective study. *Lancet* **356**, 9226, 297–302.
- Notredame, C., Higgins, D. G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 1, 205–217.
- Ortí, L., Carbajo, R. J., Pieper, U., Eswar, N., Maurer, S. M., Rai, A. K., Taylor, G., Todd, M. H., Pineda-Lucena, A., Sali, A. and Marti-Renom, M. A. (2009) A kernel for open source drug discovery in tropical diseases. *PLoS Negl Trop Dis* **3**, 4, e418.
- Oxborough, R. M., Mosha, F. W., Matowo, J., Mndeme, R., Feston, E., Hemingway, J. and Rowland, M. (2008) Mosquitoes and bednets: testing the spatial positioning of insecticide on nets and the rationale behind combination insecticide treatments. *Ann Trop Med Parasitol* **102**, 8, 717–727.
- Patel, V., Booker, M., Kramer, M., Ross, L., Celatka, C. A., Kennedy, L. M., Dvorin, J. D., Duraisingh, M. T., Sliz, P., Wirth, D. F. and Clardy, J. (2008) Identification and characterization of small molecule inhibitors of *Plasmodium falciparum* dihydroorotate dehydrogenase. *J Biol Chem* **283**, 50, 35078–35085.
- Pieper, U., Eswar, N., Webb, B. M., Eramian, D., Kelly, L., Barkan, D. T., Carter, H., Mankoo, P., Karchin, R., Marti-Renom, M. A., Davis, F. P. and Sali, A. (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **37**, Database issue, D347–D354.
- Prade, L., Jones, A. F., Boss, C., Richard-Bildstein, S., Meyer, S., Binkert, C. and Bur, D. (2005) X-ray structure of plasmepsin II complexed with a potent achiral inhibitor. *J Biol Chem* **280**, 25, 23837–23843.

- Quesada-Soriano, I., Leal, I., Casas-Solvas, J. M., Vargas-Berenguel, A., Barón, C., Ruiz-Pérez, L. M., González-Pacanoska, D. and García-Fuentes, L. (2008) Kinetic and thermodynamic characterization of dUTP hydrolysis by *Plasmodium falciparum* dUTPase. *Biochim Biophys Acta* **1784**, 9, 1347–1355.
- Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, Web Server issue, W116–W120.
- Rawlings, N. D., Barrett, A. J. and Bateman, A. (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* **40**, Database issue, D343–D350.
- Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 6, 276–277.
- Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M. and Bourne, P. E. (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* **39**, Database issue, D392–D401.
- Sadasivaiah, S., Tozan, Y. and Breman, J. G. (2007) Dichlorodiphenyltrichloroethane (DDT) for indoor residual spraying in Africa: how can it be used for malaria control? *Am J Trop Med Hyg* **77**, 6 Suppl, 249–263.
- Sajid, M. and McKerrow, J. H. (2002) Cysteine proteases of parasitic organisms. *Mol Biochem Parasitol* **120**, 1, 1–21.
- Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, Database issue, D449–D451.
- Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J. and Schomburg, D. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res* **39**, Database issue, D670–D676.

- Schomburg, I., Chang, A. and Schomburg, D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* **30**, 1, 47–49.
- Sharakhova, M. V., Hammond, M. P., Lobo, N. F., Krzywinski, J., Unger, M. F., Hillenmeyer, M. E., Bruggner, R. V., Birney, E. and Collins, F. H. (2007) Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biol* **8**, 1, R5.
- Shuto, S., Minakawa, N., Niizuma, S., Kim, H.-S., Wataya, Y. and Matsuda, A. (2002) New neplanocin analogues. 12. Alternative synthesis and antimalarial effect of (6'R)-6'-C-methylneplanocin A, a potent AdoHcy hydrolase inhibitor. *J Med Chem* **45**, 3, 748–751.
- Sidhu, A. B. S., Verdier-Pinard, D. and Fidock, D. A. (2002) Chloroquine resistance in *Plasmodium falciparum* malaria parasites conferred by *pfprt* mutations. *Science* **298**, 5591, 210–213.
- Spycher, C., Rug, M., Klonis, N., Ferguson, D. J. P., Cowman, A. F., Beck, H.-P. and Tilley, L. (2006) Genesis of and trafficking to the Maurer's clefts of *Plasmodium falciparum*-infected erythrocytes. *Mol Cell Biol* **26**, 11, 4074–4085.
- Stockwell, B. R., Haggarty, S. J. and Schreiber, S. L. (1999) High-throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications. *Chem Biol* **6**, 2, 71–83.
- Tanaka, N., Nakanishi, M., Kusakabe, Y., Shiraiwa, K., Yabe, S., Ito, Y., Kitade, Y. and Nakamura, K. T. (2004) Crystal structure of S-adenosyl-L-homocysteine hydrolase from the human malaria parasite *Plasmodium falciparum*. *J Mol Biol* **343**, 4, 1007–1017.
- Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E. and Stein, L. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* **8**, 3, R39.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. Q. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J. H., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, C., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon,

M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z. M., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W. M., Gong, F. C., Gu, Z. P., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z. X., Ketchum, K. A., Lai, Z. W., Lei, Y. D., Li, Z. Y., Li, J. Y., Liang, Y., Lin, X. Y., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B. X., Sun, J. T., Wang, Z. Y., Wang, A. H., Wang, X., Wang, J., Wei, M. H., Wides, R., Xiao, C. L., Yan, C. H., Yao, A., Ye, J., Zhan, M., Zhang, W. Q., Zhang, H. Y., Zhao, Q., Zheng, L. S., Zhong, F., Zhong, W. Y., Zhu, S. P. C., Zhao, S. Y., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H. J., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H. Y., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X. J., Lopez, J., Ma, D., Majoros, W., McDaniel, J.,

- Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M. Y., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. H. (2001) The sequence of the human genome *Science* **291**, 5507, 1304–+.
- Vértessy, B. G. and Tóth, J. (2009) Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases. *Acc Chem Res* **42**, 1, 97–106.
- Ward, P., Equinet, L., Packer, J. and Doerig, C. (2004) Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics* **5**, 79.
- Weisman, J. L., Liou, A. P., Shelat, A. A., Cohen, F. E., Guy, R. K. and DeRisi, J. L. (2006) Searching for new antimalarial therapeutics amongst known drugs. *Chem Biol Drug Des* **67**, 6, 409–416.
- Wess, J. (1997) G-protein-coupled receptors: molecular mechanisms involved in receptor activation and selectivity of G-protein recognition. *FASEB J* **11**, 5, 346–354.
- Whittingham, J. L., Leal, I., Nguyen, C., Kasinathan, G., Bell, E., Jones, A. F., Berry, C., Benito, A., Turkenburg, J. P., Dodson, E. J., Perez, L. M. R., Wilkinson, A. J., Johansson, N. G., Brun, R., Gilbert, I. H., Pacanowska, D. G. and Wilson, K. S. (2005) dUTPase as a platform for antimalarial drug design: structural basis for the selectivity of a class of nucleoside inhibitors. *Structure* **13**, 2, 329–338.
- Wu, T. Y.-H. and Ding, S. (2007) 2 - Target validation in chemogenomics. in W.Metcalf, B. and Dillon, S., editors, *Target Validation in Drug Discovery* 27 – 39 Academic Press, Burlington ISBN 978-0-12-369393-8.
- Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D. and Altman, R. B. (2004) Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome Res* **14**, 5, 917–924.
- Zhang, K. and Rathod, P. K. (2002) Divergent regulation of dihydrofolate reductase between malaria parasite and human host. *Science* **296**, 5567, 545–547.