

Homology-based *in silico* identification of putative protein-ligand interactions in the malaria parasite

by

Michał Jerzy Szolkiewicz

Submitted in partial fulfillment of requirements for the degree Magister Scientiae

in the Faculty of Natural and Agricultural Sciences

Bioinformatics and Computational Biology Unit

Department of Biochemistry

University of Pretoria

Pretoria

February 10, 2014

Declaration

I, Michał Jerzy Szolkiewicz, declare that the thesis/dissertation, which I hereby submit for the degree *Magister Scientiae* at the University of Pretoria has not been previously submitted by me for degree purposes at any other University and I take note that, if the thesis/dissertation is approved, I have to submit the additional copies, as stipulated by the relevant regulations, at least six weeks before the following graduation takes place and if I do not comply with the stipulations, the degree will not be conferred upon me.

SIGNATURE..... Date.....

Acknowledgements

- My supervisor Prof. Fourie Joubert for all his support and guidance during the course of my Msc.
- All my fellow friends at the Bioinformatics and Computational Biology Unit for all their help.
- My wife, without whom I would not have been able to finish this degree
- My adorable son, without whom this degree would have been completed 2 years prior.
- The National Research Foundation for awarding me the bursary allowing me to complete this degree.

Contents

1	Introduction	3
1.1	Malaria Endemic	3
1.1.1	Malaria	3
1.1.2	Life stages and Pathogenesis of <i>Plasmodium falciparum</i>	3
1.1.3	Anti-Malarial Drugs	5
1.1.4	Vaccine Development	8
1.2	Emergence of Resistance	9
1.3	Protein-ligand interactions	9
1.4	Physical properties of protein-ligand interactions	12
1.4.1	Measures	12
1.4.2	Interactions	12
1.4.3	Enthalpic and Entropic Contributions to Ligand-Receptor Binding . .	13
1.5	Chemogenomic approaches	13
1.5.1	Describing ligand and target space	14
1.5.1.1	Ligand space	14
1.5.1.2	Target space	17
1.5.1.3	Target-ligand space	18
1.5.2	Ligand-based chemogenomics approaches	18
1.5.2.1	Ligand-based <i>in silico</i> screening	19
1.5.3	Target-based chemogenomics	21
1.5.3.1	Sequence-based comparison	21

1.5.3.2	Domain fishing	23
1.5.3.3	Structure-based comparison	23
1.5.3.4	Target-ligand approaches	24
1.6	Ligand-based approaches	25
1.6.1	QSAR	25
1.6.1.1	Comparative Molecular Field Analysis (CoMFA)	25
1.6.1.2	Comparative Molecular Similarity Indices Analysis (CoMSIA)	26
1.6.2	Structure-based virtual screening	26
1.6.2.1	Basic requirements	27
1.6.2.2	Challenges to Docking based methods	30
1.7	ChEMBL	32
1.8	Malaria centered bio-activity data sets	34
1.8.1	GlaxoSmithKline TCAMS data set	34
1.8.1.1	A summary of the methods	34
1.8.2	Novartis-GNF Malaria Box	35
1.8.3	St Judes Children's Research Hospital Malaria data set	36
1.9	Discovery	36
1.10	Problem Statement	37
1.11	Aims	37
2	Methods	39
2.1	Overview	39
2.2	ChEMBL database integration	39
2.2.1	Chemical Data Integration	41
2.2.2	Protein Data Integration	41
2.2.3	Assay data integration	42
2.3	Protein Matching using BLAST	42
2.3.1	BLAST	42
2.3.2	The BLAST Algorithm	42

2.3.3	Application	43
2.4	Protein Matching by Domain	44
2.4.1	InterPro	44
2.4.2	InterProScan	48
2.4.3	Application	49
2.5	Clinical Trials exploration	50
2.5.1	Clinical Trial	50
2.5.2	Clinicaltrials.gov	51
2.6	Discussion	51
3	Results	53
3.1	Discovery v1.0	53
3.1.1	Primary search page:	54
3.1.2	Results of search by protein name	54
3.1.3	Results of search by ligand	55
3.2	Discovery v2.0	57
3.2.1	Chemical Searching	57
3.2.2	Protein-Ligand interactions Display	60
3.2.2.1	By Target	60
3.2.2.2	By Ligand	62
3.2.3	Clinical Trials	63
3.2.4	Advanced Search	65
3.3	Data Statistics	65
3.3.1	BLAST results	66
3.3.2	Domain results	67
3.4	Discussion	69
4	Validation	70
4.1	Case Studies	70

4.1.1	Tanimoto Distance Significance	71
4.1.2	Lactate Dehydrogenase	71
4.1.2.1	Discovery 2.0 Compounds	76
4.1.2.2	Literature Compounds	76
4.1.2.3	Remarks	77
4.1.3	Dihydroorotate dehydrogenase	77
4.1.3.1	Discovery 2.0 Compounds	78
4.1.3.2	Literature Sources	81
4.1.4	Aspartate carbamoyltransferase	82
4.1.4.1	Discovery 2.0 Compounds	82
4.1.4.2	Literature compounds	85
4.1.4.3	Remarks	85
4.1.5	Sulfadoxine	87
4.1.6	Pyrimethamine	87
4.2	Discussion	87
5	Concluding Discussion	90
	Bibliography	92

List of Figures

1.1	An overview of the <i>P. falciparum</i> life cycle within the human host	5
1.2	Enthalpic and entropic contributions	14
1.3	Examples of molecular descriptors	16
1.4	Protein representations	17
1.5	Target-Ligand descriptors	18
1.6	Ligand based target fishing summary	20
1.7	Sequence-based comparison of targets	22
1.8	Molecular Interaction Field usage	24
1.9	Docking and Scoring illustration	27
2.1	Flow Diagram of Protein-ligand interaction prediction in Discovery 2.0	40
2.2	Three basic assay designs for studying interactions	45
3.1	A screen capture of the landing page of Discovery 1.0 with the compound spermidine sketched with the Marvin sketch tool.	54
3.2	The results page of a protein view with the Ligands section selected, expanding specifically on the top Drugbank hit to indicate its matching compounds	55
3.3	The results of a ligand based search in Discovery 1.0.	56
3.4	The detail page after finding a molecule of interest within Discovery 1.0.	56
3.5	A comparison of available search types available through the ChemAxon search tool, this shows the contrasting properties of each of the search types available	58

3.6	The Chemical Search Page: Spermidine has been drawn, the SMILES string and compound name entered, only one of these is necessary to perform a search. . . .	59
3.7	Chemical Search results page	60
3.8	The Protein-Ligand interactions tab (“By target”) after selecting BLAST hit. . . .	61
3.9	The Protein-Ligand interactions tab (“By target”) after selecting Domain hit. . .	62
3.10	The Protein-Ligand interactions tab (“By ligand”) after selecting a compound. . .	63
3.11	The Clinical trials tab.	64
3.12	The advanced search feature showing a protein-ligand interactions filter.	65
3.13	A stacked bar chart indicating the E-value of the best hit per <i>Plasmodium</i> protein that has at least one BLAST hit.	66
3.14	A stacked bar chart indicating the number of BLAST hits found per <i>Plasmodium</i> protein that has at least one domain hit	67
3.15	A stacked bar chart indicating the number of shared domains of the best hit . . .	68
3.16	A stacked bar chart indicating the number of hits that have a shared domain for each <i>Plasmodium</i> protein	68
4.1	Tanimoto distance distributions of random compounds	72
4.2	The reaction catalyzed by LDH	72
4.3	Reaction catalyzed by DHOD (Patel et al., 2008).	78
4.4	Leflunomide was a compound that was in the literature set and identified as a match in Discovery 2.0	81
4.5	Reaction catalyzed by ATCase	82
4.6	The 2D structure of the molecule found in Discovery(a) and the one predicted through docking methods and found in literature(b).	86
4.7	Cross-Reference of matching ligand in Aspartate carbamoyltransferase	86
4.8	Clinical-trials showing sulfadoxine-pyrimethamine combination	88
4.9	Sulfadoxine as viewable via the Discovery2.0 interface	88
4.10	Pyrimethamine as viewable via the Discovery2.0 interface.	88

List of Tables

1.1	A summary of currently available therapeutics for malaria treatment (Dhanawat et al., 2009)	7
1.2	A summary of currently available therapeutics for malaria treatment continued.	8
1.3	Ligand descriptors	16
1.4	Bio-activity databases	19
4.1	Summary of annotation data of <i>Plasmodium</i> lactate dehydrogenase (PF13_0141), excluding protein-ligand information.	74
4.2	A Summary of the protein-ligand interactions data of lactate dehydrogenase (PF13_0141).	75
4.3	Summary of annotation data of plasmodium Dihydroorotate dehydrogenase (PFF0160c), excluding protein-ligand information.	79
4.4	A Summary of the protein-ligand-interactions data of lactate dehydrogenase (PF13_0141).	80
4.5	Summary of annotation data of plasmodium sspartate carbamoyltransferase (MAL13P1.221), excluding protein-ligand information.	83
4.6	A summary of the protein-ligand-interactions data of aspartate carbamoyltransferase (MAL13P1.221).	84

Nomenclature

ADMET	An abbreviation term referring to absorption, distribution, metabolism, excretion and toxicity properties of chemical compounds
erythrocytic stage	pertaining to, characterized by, or of the nature of erythrocytes.
exoerythrocytic stage	The developmental stage of the malaria parasite in liver parenchyma cells of the vertebrate host before the red blood cells become infected.
gametocytes	a eukaryotic germ cell that divides by mitosis into other gametocytes
GPCR	G protein-coupled receptors
HTS	High Throughput Screening
IDC	Intraerythrocytic development cycle- A stage in malaria after the parasite has penetrated the host red blood cell.
IFP	Interaction Finger Print
LDH	Lactate Dehydrogenase
merozoite	Add Description
MIF	Molecular Interaction Field
oocyst	the encysted or encapsulated ookinete in the wall of a mosquito's stomach; also, the analogous stage in the development of any sporozoan.

ookinete	the fertilized form of the malarial parasite in a mosquito's body, formed by fertilization of a macrogamete by a microgamete and developing into an oocyst.
Plasmodium falciparum	The species and genus name given to one of the malaria causing parasites.
QSAR	Quantitative structure–activity relationship
schizont	A sporozoan cell that reproduces by schizogony, producing a varied number of daughter trophozoites or merozoites.
TCAMs	Tres Cantos Antimalarial dataset
trophozoite	the active, motile feeding stage of a sporozoan parasite.

Abstract

Malaria is still one of the most prolific communicable diseases in the world with more than 200 million infections annually, its greatest effect is felt in the poor nations with-in sub-saharan Africa and south-east Asia. It is especially fatal for women and children where out of the 660 000 fatalities in 2010, 86% were below the age of 5.

In the past decade the global fatality rate due to malaria has been significantly reduced, primarily due to proliferation of vector control using treated nets and indoor residual spraying of DDT. There have, however, been few innovations in anti-malarial therapeutics and with the threat of the spread of drug resistant strains a need still exists to develop novel drugs to combat malaria infections. One of the major hinderances to drug development is the huge cost of the drug development process, where candidate failures late in development are extremely costly. This is where post-genomic information has the potential of adding great value. By using all available data pertaining to a disease, one gains higher discerning power to select good drug candidates and identify risks early in development before serious investments are made. This need provided the motivation for the development of Discovery; a tool to aid in the identification of protein targets and viable lead compounds for the treatment of malaria. Discovery was developed at the University of Pretoria to be a platform for a large spectrum of biological data focused on the malaria causing *Plasmodium* parasite. It conglomerates various data types into a web-based interface that allows searching using logical filters or by using protein or chemical start points. In 2010 it was decided to rebuild Discovery to improve it's functionality and optimize query times. Also, since its inception various new datasources became available specifically related to bio-active molecules, these include the

ChEMBL database and TCAMS dataset of bio-active molecules and the focus of this project was the integration of said datasets into Discovery. Large quantities of high quality bio-activity data have never been available in the public domain and this has opened up the opportunity to gain even greater insight into the activity of chemical compounds in malaria. Due to conserved structural/functional similarities of proteins between different species it is possible to derive predictions about a malaria protein or a chemicals activity in malaria due to experiments carried out on other organisms. These comparisons can be leveraged to highlight potential new compounds that were previously not considered or prevent wasting resources persuing potential compounds that pose threats of toxicity to humans. This project has resulted in a web based system that allows one to search through the chemical space of the malaria parasite. Allowing them to view sets of predicted protein-ligand interactions for a given protein based on that proteins similarity to those existing in the bio-active molecule databases.

Chapter 1

Introduction

1.1 Malaria Endemic

1.1.1 Malaria

Plasmodium falciparum (*P. falciparum*) is one of five malaria causing parasites from the genus *Plasmodium* that are able to infect humans. *P. falciparum* is responsible for the majority of deaths caused by malaria in the world (Gardner et al., 2002). According to the World Health Organisation (WHO) malaria report 2012 (Organization, WHO), deaths caused by malaria are estimated at about 655 000 but could be as high as 900 000 due to poor surveillance in many countries. The majority of deaths occur in Africa and are children and females. Malaria is endemic in poorer regions in the world and is a major cause of poverty in these nations, to the extent that the cure for malaria is being seen as a tool to alleviate the poverty problem in these countries (Teklehaimanot and Mejia, 2008).

1.1.2 Life stages and Pathogenesis of *Plasmodium falciparum*

The *Plasmodium* parasite is transmitted through the bite of the female mosquito, *Anopheles gambiae* which acts as the primary host of the parasite. The mosquito bites an individual infected with malaria and the parasite gametocytes are taken up into the gut of the mosquito where they develop into male and female gametes and fuse together in the gut to form a

ookinete that penetrates the cell wall of the gut and forms a oocyst. When the oocyst develops and ruptures, sporozoites are released and these migrate to the mosquito's salivary glands, ready to infect another human. Upon the next bite the parasite is transferred into its new host through the saliva (Bledsoe, 2005).

Malaria in humans develops in two main phases, exoerythrocytic and erythrocytic phases, meaning outside of red blood cells and internally within red blood cells respectively. Immediately after infection of the human host the Plasmodium sporozoites migrate to the liver and infect the hepatocytes. They then multiply asexually and differentiate into thousands of merozoites which rupture the host cells, and move to infect red blood cells, this is the start of the intraerythrocytic development cycle (IDC). In this stage the parasite goes through various identifiable stages in the sequence of the ring stage, the trophozoite also known as the feeding stage, schizont also known as the reproduction stage and return to merozoite (which occurs prior to entering the erythrocyte). Figure 1.1 gives an illustration of the cycle that takes place within the red blood cells. During the IDC the parasite expresses a large variety of proteins to allow it to proceed through its various stages, this makes the IDC the focus of research efforts because of the large number of key cycles that need to proceed timeously for the parasite to survive, leaving many opportunities to disrupt any of these stages as to is the current focus of most anti-malarial drugs (Bozdech et al., 2003).

Pathogenesis due to malaria is caused by the parasite entering the erythrocytic phase where it infects a red blood cell and reproduces asexually, then the new cells periodically break out and infect new blood cells. A symptom of this is the patient suffering periodic waves of fever due to the release of merozoites and reinfection of the new cells. Malaria causes severe anemia as well as the potential to cause cerebral malaria which is a characteristic of *P. falciparum* where the infected red blood cells can pass through the blood brain barrier which can lead to the patient suffering from a coma and later death. The *P. falciparum* cells avoid detection from the human immune system because most of the life cycle in the

human host is spent within liver and red blood cells where it is relatively invisible to immune surveillance. Infected blood cells that remain in circulation can be destroyed by the spleen, but the parasite presents adhesive proteins on the membrane of the red blood cell causing it to stick to the walls of blood vessels and these cells could then cause blockage in venules, the blocking of these venules could cause the symptoms of cerebral malaria (Miller et al., 1994) .

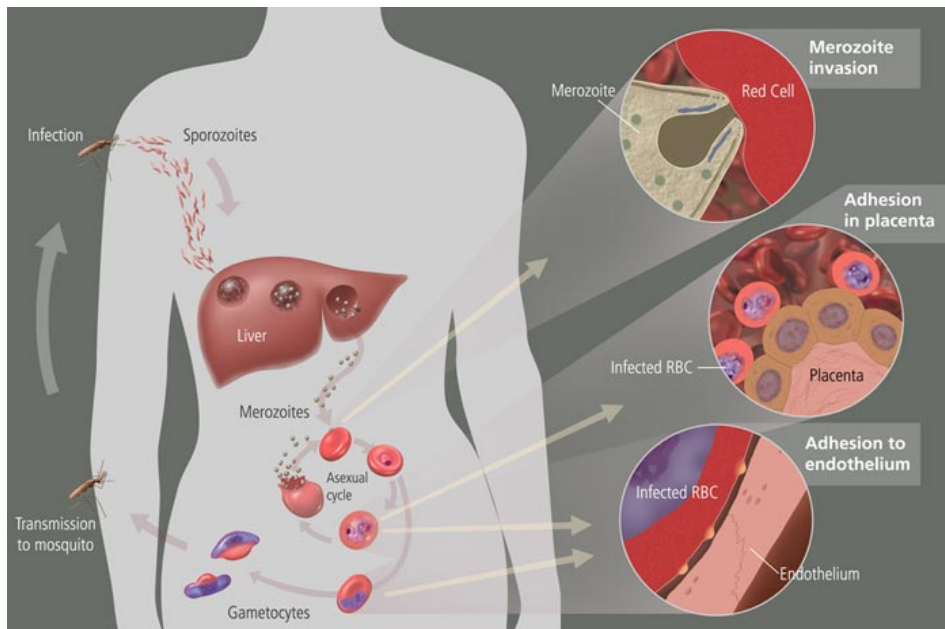


Figure 1.1: An overview of the *P. falciparum* life cycle within the human host. (Image courtesy of the Medical Arts and Photography Branch, NIH). The parasite is transferred between individuals infected with malaria through the mosquito, sporozoites then move to the liver hepatocytes where they develop further and multiply into merozoites, the hepatocytes rupture and release the merozoites into the bloodstream where they infect red blood cells and enter the intraerythrocytic development cycle and eventually develop into gametocytes which can then be taken up by another mosquito.

1.1.3 Anti-Malarial Drugs

Current drug treatments available are in two categories, firstly in the form of prophylactics where an individual is treated in order to prevent infection by the parasite. The prophylactics are further subdivided into suppressive which can only work once the parasite has reached the erythrocytic stage and causal prophylactics that can target the liver stage of malaria development. Suppressive prophylactics therefore have no effect until the liver stage is complete, these drugs include chloroquine, proguanil, mefloquine, and doxycycline. The second

category is the treatment of individuals with suspected or confirmed infection, the drugs vary widely in activity and mechanism but one example is pyrimethamine which prevents DNA synthesis and can act on schizonts in both hepatic and erythrocytic phases (White, 2004).

To summarise the current drugs available for the chemotherapeutic treatment of malaria Tables 1.1 and 1.2 (Dhanawat et al., 2009) show the year in which it was first used, its class, its source (either synthetic, semi synthetic or natural), remarks about the drugs and their targets (Kappe et al., 2010).

S. No	Year	Drug	Class	Source	Remarks	Targets
1	1820	Quinine	Alkaloid	C	- Hemolytic anemia in patients with G6PD deficiency	- Inhibition of hemozoin biocrystallization, thus facilitating an aggregation of cytotoxic heme
2	1926	Pamaquine	8-aminoquinoline	A	- Hemolytic anemia in patients with G6PD deficiency	- Generating reactive oxygen species or interfering with the electron transport in the parasite.
3	1931	Quinacrine	Acridine	A	- Also used in systemic lupus erythematosus	Act against the protozoan's cell membrane.
4	1934	Chloroquine	4-aminoquinoline	A	- Adversely effect immune system - 1950 resistance appears	- High alkaline pH in food vacuoles of the parasite
5	1935	Sulfadoxine	Sulphonamide	A	- Serious (possibly fatal) allergic reactions	- Inhibits dihydropteroate synthase (DHPS, EC 2.5.1.15) - Key enzymes in the biosynthesis of folate
6	1945	Proguanil	Biguanide	A	- Mouth ulcers, skin rashes, reversible hair loss, Severe kidney impairment	- Inhibiting the enzyme, dihydrofolate reductase
7	1950	Primaquine	8-aminoquinoline	A	- Effective against the gametocytes - Hemolytic anemia in patients with G6PD deficiency	- By generating reactive oxygen species or by interfering with the electron transport in the parasite
8	1951	Cycloguanil	Biguanide	A	- Active metabolite of proguanil	- Acts specifically on <i>P. falciparum</i> DHFR (EC 1.5.1.3)
9	1970	Amodiaquine	4-aminoquinoline	A	- Agranulocytes and hepatic disorders	- Inhibit heme polymerase activity - Accumulation of free heme
10	1970	Pyrimethamine	Diamino pyrimidine	A	- Adversely effect immune system - Bone marrow depressants - Megaloblastic anemia if used with other folate antagonists - Hepatic function impairment	- Folic acid antagonist
11	1970	Pyronaridine	9-anilinoacridine	A	- Effective against chloroquine-resistant parasites	- Blood schizontocidal
12	1970	Lumefantrine (benflumetol)	Fluorenemethanols	A	- Artemisinin-based combination therapy	Forming toxic complexes with ferriprophyrin
13	1971	Mefloquine	Arylamino alcohol	A	- Negative inotropic effects - Prevention and treatment of malaria	- Blood schizonticide. Its exact mechanism of action is not known.

Table 1.1: A summary of currently available therapeutics for malaria treatment (Dhanawat et al., 2009)

S. No	Year	Drug	Class	Source	Remarks	Targets
14	1971	Artemisinin	sesquiterpene lactones	C	<ul style="list-style-type: none"> - Water-soluble (most active and the least toxic) - lipid-soluble (longest life but also the most toxic) - Oil soluble 	<ul style="list-style-type: none"> - Inhibiting a <i>P. falciparum</i>-encoded sarcoplasmic-endoplasmic reticulum calcium ATPase
		Artesunate		B		
		Artemether				
		Arteether				
		Dihydroartemisinin				
		Artelinic acid				
Artenimol						
15	1980s	Halofantrine	9-phenanthrine methanol	A	<ul style="list-style-type: none"> - Cardiac arrhythmia - Short half life 	<ul style="list-style-type: none"> - Forming toxic complexes with ferritoporphyrin IX that damage the membrane of the parasite
16	1980s	Atovaquone	1,4 naphthoquinone	A	<ul style="list-style-type: none"> - Only available as a fixed preparation with proguanil (Malarone) 	<ul style="list-style-type: none"> - Site of action cytochrome bc1 complex (Complex III) - Inhibition of mitochondrial electron transport
17	2007	Tafenoquine (WR-238605) (Under development)	8-aminoquinoline	A	<ul style="list-style-type: none"> - long half-life - Use for Malaria Prophylaxis - Hemolysis in people with G-6-PD deficiency 	<ul style="list-style-type: none"> - Sporontocidal and gametocytocidal activity

A=Synthetic, B=Semisynthetic, C=Natural.

Table 1.2: A summary of currently available therapeutics for malaria treatment continued.

1.1.4 Vaccine Development

Out of all the vaccines in development RTS,S/AS01 is the most advanced and is at least 5 years ahead of any other candidate (Schwartz et al., 2012). The first results of the large scale Phase 3 clinical trial that was conducted on RTS,S/AS01 were released in 2011. The trial showed that RTS,S/AS01 has made some progress in the development of an effective prophylactic but unfortunately has not reached the effectiveness that would be needed for it to be viewed as a solution to the malaria problem, the vaccine was able to reduce infection rate by between 45 and 55% (Lell et al., 2011).

1.2 Emergence of Resistance

Strains of *Plasmodium* have emerged with resistance to many of the current therapeutic drugs. The first occurrence of resistance was noted in 1957 to chloroquine, in this case the parasite developed an efflux mechanism that expels chloroquine from the parasite before a high enough level is reached within the parasite (Krogstad et al., 1987). Resistance is also thought to occur as a result of point mutations. As an example *Plasmodium* developed resistance to antifolate combination drugs, the most common being sulfadoxine and pyrimethamine which is thought to have occurred through two point mutations allowing blockages of enzymes involved in folate synthesis. The reason for the parasite adapting and developing resistance to these drugs can be one or a combination of many reasons, an example being due to the genetic flexibility and immunogenic complexity of *P. falciparum*, and due to a lack of control and distribution of available drugs, for example institutes dispensing dilute versions of the drugs or the prescriptions not being completed by the patients allow for the parasite to be exposed to non-lethal doses of the drugs which allow for resistance to develop in the parasite (Ridley, 2002).

1.3 Protein-ligand interactions

This project deals primarily with protein-ligand interactions and their identification. The following sections will provide some background into understanding what protein-ligand interactions are and which methods researchers employ when searching for them.

Predicting protein-ligand interactions is a vital step into deciphering many biological processes, they are essential for understanding processes like signal transduction; and they play a vital role in drug discovery. The theory behind an interaction between a ligand and a protein historically followed the E. Fischer “lock and key” model. The “lock” represents the protein which interacts with a “key” which represents the ligand, the key needs to fit correctly into the keyhole (binding site) in the correct orientation in order to exert an effect (Koshland,

1995). This approach however is an oversimplification of the problem and the “hand-in-glove” analogy was then developed to better represent the dynamics in a protein-ligand interaction, that is adding that the protein and the ligand are both flexible units and during the course of the interaction process both the ligand and the protein adjust their conformations to create the “best fit” (Jorgensen, 1991). This conformation changing and fitting is known as an “induced fit”.

The usual methods for identifying novel protein-ligand interactions can be classified into two groups; namely ligand-based and structure or docking based approaches. Ligand based approaches typically compare a ligand that is already known to interact or inhibit a protein to unknown ligands in search of new interacting molecules. These methods usually use machine learning algorithms (Butina et al., 2002). Structure based approaches attempt to predict how a candidate ligand will interact with a protein by using the 3D structure of the protein to attempt to fit the ligand into its active site (Halperin et al., 2002).

In order to successfully apply a ligand-based approach, a researcher needs to have knowledge of enough interacting compounds of a protein to be able to generate an accurate prediction. If no ligands are known to interact with a protein it is then still feasible to use a structure-based approach, however, if the 3D structure of the protein target is not known and cannot be derived then none of the classic methods can be applied. Both these methodologies require that a researcher look at a single protein target independently of other proteins, thus a new concept called chemogenomics was developed (Caron et al., 2001). Chemogenomics aims to mine the entire chemical space of an organism implying that a set of all small molecules be mapped to the biological space referring to a set of all proteins or at least protein families. A reasonable assumption to motivate chemogenomics is that some classes of molecules can bind similar proteins, which suggest that if we have knowledge of some ligands for protein targets, it is possible to find ligands for similar targets (Rognan, 2007).

When a researcher decides to begin a drug discovery project and wants to predict protein-ligand interactions the most important consideration when deciding on an approach to use is the type of data available, if any. It is possible to begin a drug discovery study in a case where a researcher has no ligand data or protein structure by performing a high throughput screen (HTS).

HTS has come into existence due to the advances in robotics and computational techniques. It allows for millions of chemical or pharmacological experiments to be conducted very rapidly using assays where a substance be it proteins, or whole cells are exposed to a large variety of chemical compounds and their phenotypic effect is measured (Sundberg, 2000). This allows for a large amount of data to be produced and chemical start points to be identified for further study. A good example of this is the TCAMS antimalarial dataset (discussed in 1.8.1) developed by GlaxoSmithKline. This data-set was released into the public domain where a HTS was performed on *Plasmodium* grown in culture and measured the level of inhibition of 2 million compounds which led to a library of 13533 compounds found to inhibit malaria growth (Gamo et al., 2010).

If a researcher has ligand information but no targets it is possible to perform pharmacophore screening methods and 3-D QSAR to try to find the targets of the ligands. When a researcher has a protein structure of a potential drug target but does not have any ligand data the most likely approach is to perform docking studies against large libraries of compounds and another option is to perform a *de novo* design and build a chemical structure from analyzing the binding site of a protein and predicting what an interacting ligand will look like. And if a researcher has a protein structure and interacting ligands a structure based drug design method can be used to perform lead optimization.

1.4 Physical properties of protein-ligand interactions

1.4.1 Measures

It is firstly important to note that interactions are measured by a value called a dissociation constant which is used to measure affinity of ligands with enzymes but relates to all interactions in the same way

K_i is measured by taking the protein and compound in solution and allowing the solution to reach equilibrium, then K_i is equal to the concentration of the ligand multiplied by the concentration of the protein divided by the concentration of the protein-ligand-complex as shown in this formula:

$$K_i = \frac{[PROTEIN][LIGAND]}{[PROTEIN-LIGAND-COMPLEX]}$$

Another measure commonly used is IC_{50} which is the concentration of the ligand that reduces the activity of the protein to 50%

1.4.2 Interactions

Different kinds of interactions exist in a protein-ligand interaction in many cases these interactions occur simultaneously with different degrees of importance for the interaction:

1. Hydrogen-bonds : Hydrogen (proton) donors and acceptors.
2. Ionic interactions : Depend on (permanent or induced) ionic charges of ligands and proteins. Induced charges are influenced by the pH, pKa and pKB, electrostatic attraction. These are very strong within distances of around 3Å.
3. Metal ion complexes : Many enzymes contain metal ions, which interact with the protein via charge induced interactions or via specific molecules, e.g. thiol groups (R-SH).

4. Hydrophobic interactions : Lipophilic moieties of the ligand interact with non-polar amino acids of the protein, the energy benefit results mainly from water replacement out of hydrophobic binding pockets.

1.4.3 Enthalpic and Entropic Contributions to Ligand-Receptor Binding

Figure 1.2 illustrates the enthalpic and entropic contributions to the interaction. In summary after a ligand comes into contact with its target before an interaction can take place, both the receptor and the ligand need to be stripped of their solvent shells; the water molecules that remain form bridging hydrogen bonds; additional ions may be required for binding; the receptor might need to change shape; and the ligands rotatable bonds need to be arranged. It is important to state that the stability of the protein-ligand complex has a large influence on the bio-activity of that compound and the energy gained during the formation of the protein-ligand complex is the energy responsible for the stability of the complex. It is generally seen that more lipophilic ligands have greater affinities, this is because these compounds allow for more hydrogen bonds to form in the solvent and easily moves from the solvent to bind to hydrophobic receptor surfaces. Hydrophilic areas are however seen to increase instability of the complex, due to the water molecules that surround the ligand that interrupt the interaction, and are also seen to have discriminatory properties which contribute highly to selectivity of the binding site (A.Bender and R.C.Glen, 2004).

1.5 Chemogenomic approaches

The term chemogenomics has been defined as the discovery and description of all possible drugs to all possible drug targets (Rognan, 2007). It is closely related to the concept of chemical genetics and chemical genomics in the approaches used to investigate the relationship between biological systems with small molecules. The difference between these approaches is that chemogenomics tries to emphasize the inherent relationships between targets and

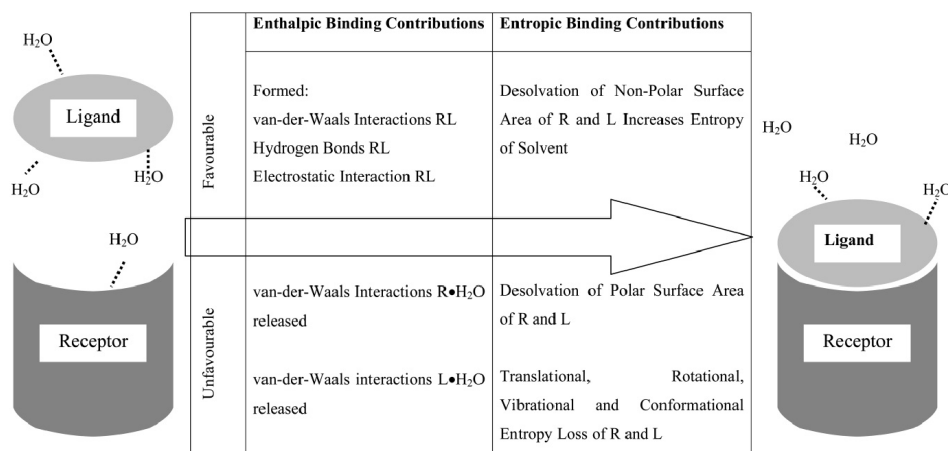


Figure 1.2: A summary of enthalpic and entropic contributions to ligand receptor binding (A. Bender and R.C. Glen, 2004).

small molecules while chemical genomics and chemical genetics emphasize the effect of a small molecule on the biological system. In recent years there has been a large increase in the amount of small-molecule bio-activity information available publicly in electronic form with the ability to process this data has become very rapid. Thus with all this information available a researcher is no longer limited to considering shared features of antagonists to one receptor but can now view characteristic features of antagonists of all known receptors and find relations between them (Caron et al., 2001).

1.5.1 Describing ligand and target space

1.5.1.1 Ligand space

Chemogenomics assumes two basic principles, the first is that compounds that are similar to one another will share targets and second that targets sharing similar ligands should share similar patterns i.e. binding sites. Thus to complete a data set using a chemogenomic approach implies that all targets that do not have ligands associated to them be linked to the nearest neighbour target that has ligands. Similarly, all ligands that do not have a target should be linked to the nearest neighbour ligand that has a target. The question here is not necessarily if these assumptions are correct but rather what is the best way to measure the distance between two targets or two ligands (Rognan, 2007).

In order to correctly identify similar compounds one needs to select appropriate descriptors. Descriptors are often classified into their dimensionality ranging between 1-D and 3-D descriptors and are summarized in Table 1.3 and Figure 1.3 (A.Bender and R.C.Glen, 2004).

The 1-D descriptors are easy to and quick to compute because they can be derived from the chemical formula, they can be used also to compute a number of additional properties for instance ADMET (absorption, distribution, metabolism, excretion and toxicity properties) and even help classify ligands as drugs or non-drugs (Sadowski and Kubinyi, 1998). They are commonly used to do very rapid comparisons and the most common descriptor is a SMILE (Simplified Molecular Input Line Entry String). The vast majority of chemical descriptors used are 2-D topological descriptors or sketches of the structure, where a connectivity table is generated that can represent both atomic and bond properties. These can then be used to search for substructure and also cluster compounds into subfamilies. The other form of 2-D descriptors is fingerprint-based methods, where a structure is coded into a bit string of 1's and 0's that represent the atoms, fragments, rings and substructures. Fingerprints have been found to often be the most appropriate method of comparison due to its speed and effectiveness (Willett, 2006).

3-D descriptors include atomic co-ordinates, 3-D pharmacophores, shapes, potentials and fields. These can be compared by either aligning molecules on the same Cartesian plane or by converting the 3-D information into a bit string which is much easier and quicker than attempting to compare the structures. Most similarity measures move to simplify the comparison to single indices, of which the most commonly used is the Tanimoto coefficient. The Tanimoto coefficient is described below.

$$Tc = \frac{c}{a+b-c}$$

where a is the number of bits in compound A , b is the number of bits in compound B and c is the number of bits in both A and B . The coefficient thus ranges between 0 and 1 where 0 represents 2 completely different compounds and 1 represents identical compounds.

Dimension	Nature	Examples
1-D	Global	Molecular weight, atom and bond counts (for example, number of H-bond donors, number of rings), polar surface area, polarizability, $\log P$
2-D	Topological	Topological and connectivity indices, fragments, substructures (for example, maximum common substructures), topological fingerprints (for example, structural keys)
3-D	Conformational	n -points pharmacophore, shape, field, spectra and fingerprints

Table 1.3: Examples of the different ligand descriptors (Rognan, 2007).

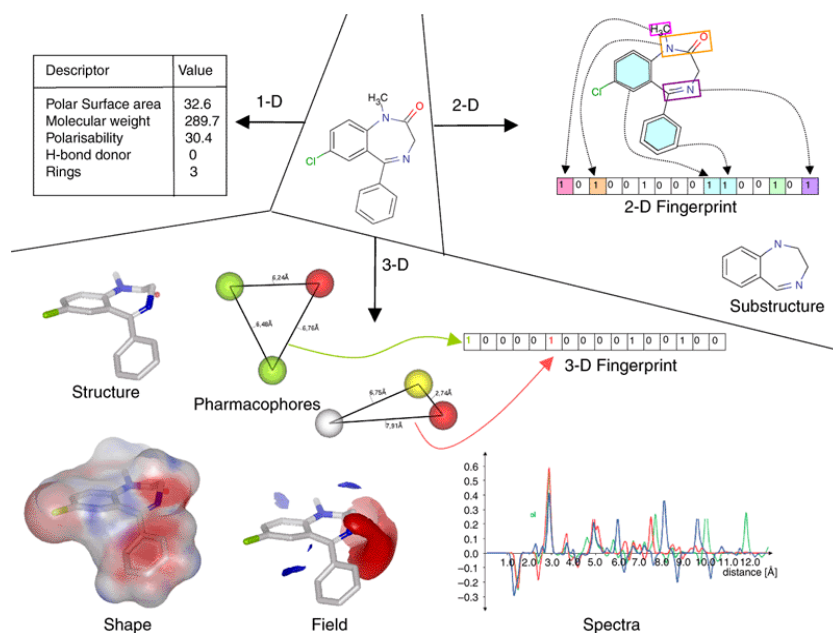


Figure 1.3: Examples of the various molecular descriptors, the 2D chemical sketch can be represented in various ways to allow searching against its properties. There are specific uses for all of these descriptors and each has pros and cons. (Rognan, 2007)

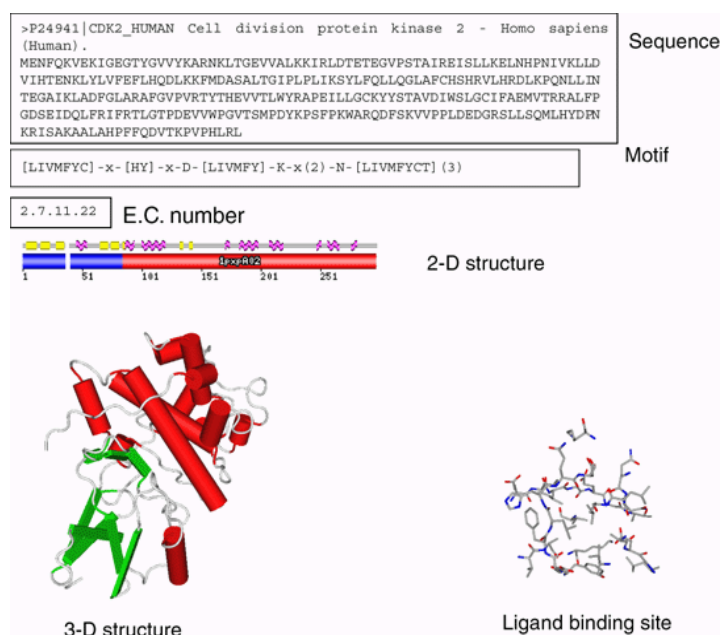


Figure 1.4: Representations of a protein using 1-D to 3-D properties (Rognan, 2007).

1.5.1.2 Target space

Proteins are most commonly classified by their sequence (1-D) and structure (3-D), at the sequence level a researcher has the ability to reliably cluster targets into families for example kinases or GPCRs. Protein sequences, even in the same family, vary in length to a large degree and that makes their alignment more difficult especially when having to deal with large insertions and deletions, so researchers often put focus on finding common motifs within sequences (Attwood et al., 2003) which are collections of residues that are specific for a protein family. The structural organization of a protein is also a very important feature to consider as it greatly effects the activity of the protein, structure can be represented in 2-D where the α -helices, β -sheets, coils and random structures are mapped to the sequence and in 3-D where a structures atomic co-ordinates are identified using methods like x-ray crystallography and NMR. The focus in most drug discovery studies is the binding site of a protein structure as it is what determines a protein's activity with a certain ligand.

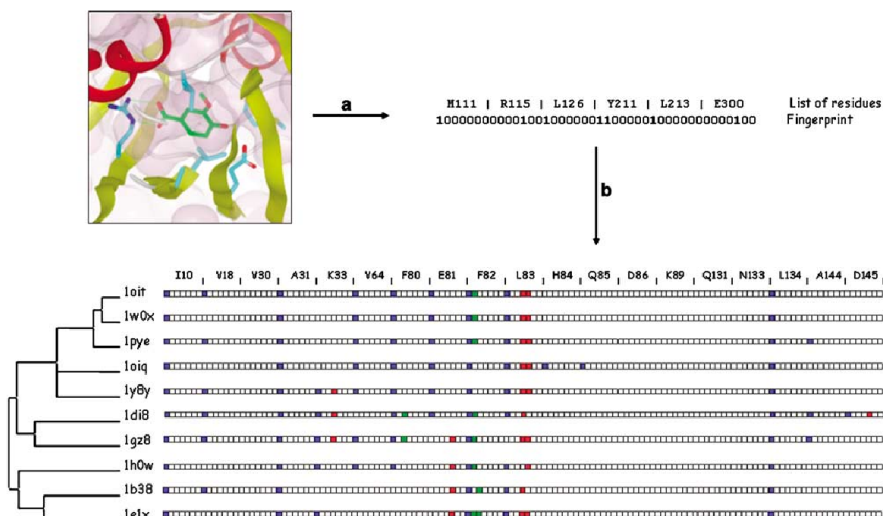


Figure 1.5: (a) Deriving and (b) comparing protein–ligand complexes by molecular interaction fingerprints. ‘0’ and ‘1’ digits are replaced by color-coded squares for the ease of comparison (blue, hydrophobic interactions; green, aromatic interactions; red, hydrogen bonds) (Rognan, 2007).

1.5.1.3 Target-ligand space

It is also possible to describe protein-ligand interactions for comparative purposes using interactions that have been experimentally evaluated and have either affinity data or structural information. The drawback to this kind of data is that it is very scarce due to the scale of the experiments to generate this type of data. A descriptor of some interest in this area is a structural IFP (interaction fingerprint), illustrated in Figure 1.5 which converts atomic co-ordinates into a bit string similar to a chemical fingerprint but carrying information on the binding site of each residue and the type of molecular interaction taking place at that position. It is then possible to compare series of complexes where one protein is matched to a series of ligands or one ligand is matched to a series of proteins (Rognan, 2007).

1.5.2 Ligand-based chemogenomics approaches

This approach attempts to pool together targets at a family or subfamily level and build a model for the ligands that will interact with that family of targets (Rognan, 2007). The basic assumption behind this approach is that if a unknown ligand is structurally similar enough to an already biologically annotated ligand then there is a better chance that they will exhibit

Databases	Data	Web address
ChEMBL [21]	>700k small molecules with >2.7 million bioactivity data points	http://www.ebi.ac.uk/chembl
DrugBank [150]	4800 drug entries including: >1350 FDA-approved drugs and 123 FDA-approved biologics	http://www.drugbank.ca/
ChemBank [24]	Information on hundreds of thousands of small molecules and hundreds of biomedical assays	http://chembank.broadinstitute.org/
Comparative Toxicogenomics Database (CTD) [151]	6000 compounds, 1.4 million chemical-gene-disease data points	http://ctd.mdibl.org/
SuperTarget [152]	1500 drugs, 2500 targets proteins and 7300 drug-target interactions	http://bioinf-tomcat.charite.de
MATADOR [152]	Manually annotated compounds from the SuperTarget database	http://matador.embl.de/
Therapeutic Target Database (TTD) [153]	Contains 1906 targets, including 358 successful, 251 clinical trial, 43 discontinued and 1254 research targets, and 5124 drugs, including 1511 approved, 1118 clinical trial and 2331 experimental drugs	http://xin.cz3.nus.edu.sg
PubChem	250k compounds, 2500 bioassays	http://pubchem.ncbi.nlm.nih.gov/
BioActivity [23]	>271k compounds, >620k binding affinities against 5526 protein targets	http://www.bindingdb.org
BindingDB [39]	Contains 1492 approved drugs and 1664 unique protein targets	http://www.ebi.ac.uk
DrugPort (EBI)	Contains 1207 entries covering 841 known and potential drug targets with structures from the Protein Data Bank (PDB)	http://www.dddc.ac.cn/pdtd/
Potential Drug Target Database (PDTD) [80]	From text mining >20 mio. Publications, includes >10 mio. Proteins, >25k compounds, as well as drug side effects and various other data; includes protein network visualization	http://bioinformatics.charite.de/promiscuous
Promiscuous [31]	Toxin and Toxin-Target Database; >2900 toxins, 1300 toxin targets and >33,000 toxin-target associations	http://www.t3db.org/
T3DB [154]	RDF-based database integrating chemical, biological and phenotypic data	http://cheminfv.informatics.indiana.edu:8080/
Chem2Bio2RDF [30]		

Table 1.4: Overview of the major public bio-activity data resources available at present (Koutsoukas et al., 2011).

similar biological properties. It is then vital to have databases of biologically annotated ligands before it is possible to perform this kind of study. Fortunately the number of these databases that are becoming publicly available is increasing at a rapid rate. A summary of these databases can be seen in Table 1.4, it is important to note that these databases are biased to pharmaceutically important target families such as GPCRs and kinases which limits the potential of identifying interactors that fall outside of those well defined families. These biologically annotated databases are a primary source of potential new biological mechanisms that can be exploited for drug discovery and development to treat diseases.

1.5.2.1 Ligand-based *in silico* screening

A ligand-based *in silico* screening approach to target-fishing (identifying targets for unannotated chemical ligands) involves 3 basic components irrespective of which method employed (Cases et al., 2005). The first is to have a set of reference compounds which have either 2-D or 3-D descriptors, secondly a procedure setup to screen by either QSAR, machine learning

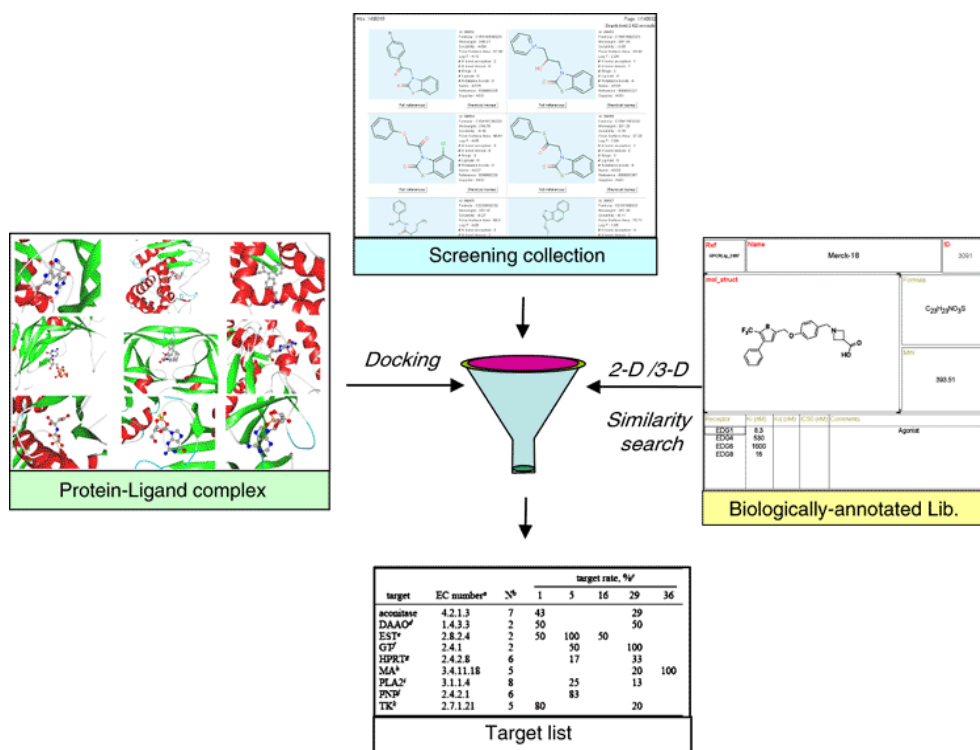


Figure 1.6: Ligand based target fishing summary. This illustrates the data required to perform a ligand based screening (Rognan, 2007).

or pharmacophore searches and lastly a screening collection to identify new molecules likely to share a target with the reference compounds. The process followed is to first categorize the compounds from the training set by their protein target without providing a binding site or describing the type of interaction taking place (i.e. whether it is an antagonist or agonist), which leaves the possibility that the machine learning algorithms could generate false rules by using incorrect data. To overcome these potential errors 3-D approaches such as pharmacophore modeling can be used (to be discussed in detail later)(Steindl et al., 2006). An illustration of the screening process is shown in Figure 1.6.

This technique was successfully applied by Novartis using a Bayesian statistics-based machine learning algorithm (Xia et al., 2004), where they predicted target profiles of compounds from the Wombat database. In their approach they created a Bayesian model for each target to distinguish between active and inactive compounds. They calculated the probability that a compound will be active against each target and selected the most likely target. This method was found to correctly predict the target 77% of the time when the training set was

from the Wombat database and the test set was from the MDDR database (Nidhi et al., 2006). When assessing other 2-D and 3-D descriptors for the same application the 2-D descriptors were found to have more predictive power than the 3-D pharmacophore approach but 3-D descriptors are more useful when considering molecules that show little similarity to the training set compounds.

1.5.3 Target-based chemogenomics

A large amount of potential value exists in comparing targets from the same family, especially those with structural data in order to perform proteome-wide comparative modeling of targets with unknown structure. Target-based chemogenomic approaches are divided into two categories, one being based on sequence information the other structure-based.

1.5.3.1 Sequence-based comparison

A sequence-based approach in principle can be used for any protein family provided that a multiple sequence alignment of the possible targets is possible. A sequence-based approach is generally applied when there is too little structural information available to be useful. GPCR's make a good example for the value of this approach because of the known importance of GPCR's in drug design and the limited number of GPCR structures available (Frimurer et al., 2005).

Once an alignment is performed on all the sequences it is possible to identify key residues that map to the binding site of most ligands. These residues can then be extracted and combined into an ungapped sequence which is used to create a distance matrix based on sequence identity, sequence similarity and physio-chemical properties, called a cavity-tree, illustrated in Fig 1.7 (Surgand et al., 2006). A study performed by Surgand *et al* clustered 372 human GPCRs using this strategy and interestingly found that it produces an identical tree to one based on the full sequence indicating that only a few residues are required to compare targets within the same family.

The cavity trees can be applied using a simple principle called target-hopping where

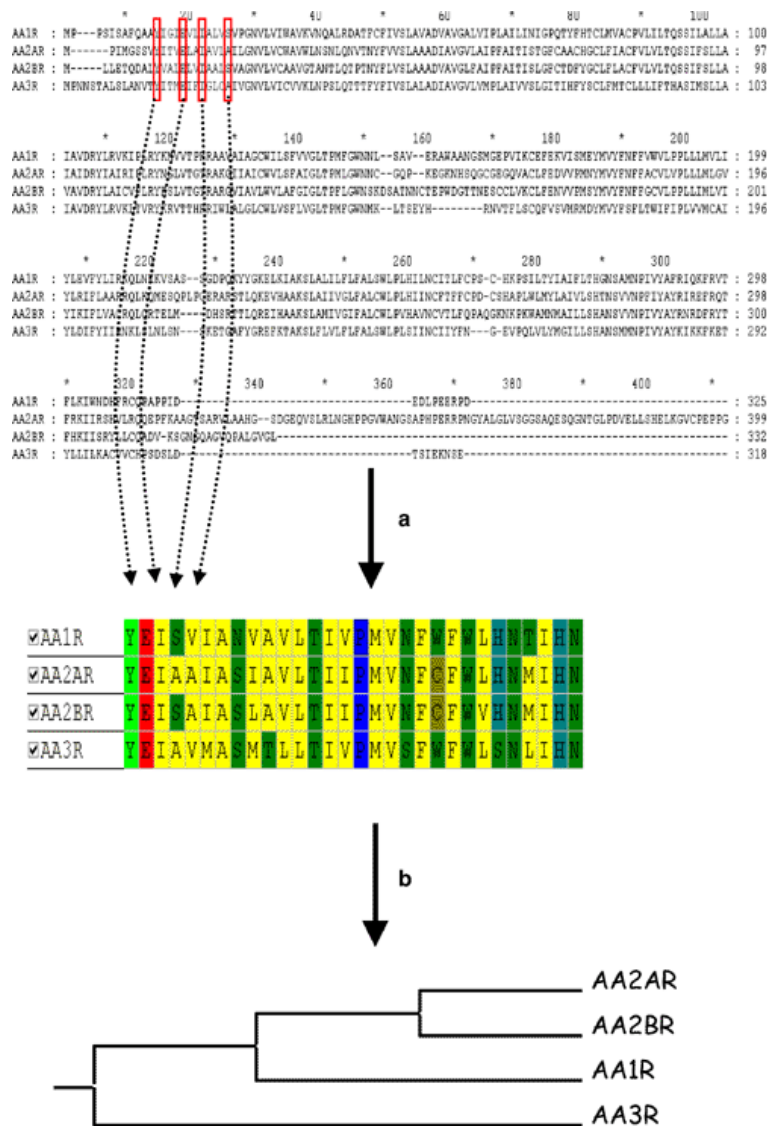


Figure 1.7: Sequence-based comparison of targets (a) Selection of key cavity-lining residues and (b) Clustering according to residue conservation (Surgand et al., 2006).

predicting new target ligands can be done by identifying known ligands of a similar receptor. As an example Frimurer *et al* used a GPCR cavity-based tree CRTH2 receptor antagonists were identified from known angiotensin II type 1 receptor antagonists (Frimurer et al., 2005).

1.5.3.2 Domain fishing

A shortcoming of predicting targets for ligands is the restrictions to the size of the training set and targets can only be confidently predicted if they have orthologs with known ligands in the database. Thus a domain fishing approach was developed by Bender and Glen (2004), where they built protein domain-based models to predict interactions with, the assumption being that similar ligands are not only likely to bind to the same target but also to the same protein folds and amino acid sequences that occur in other proteins. This approach greatly widens the net so to speak of what interactions can be predicted so it is possible to predict interactions to targets that are very distant to those in the known interactions database. This approach cannot make definitive predictions due to the complexity of protein structure but the ability to expand the scope of the targets list outweighs the negative of not being able to make definitive predictions (A.Bender et al., 2007).

1.5.3.3 Structure-based comparison

If a target family has enough good structural templates it is possible to perform structure-based comparisons. Usually only ligand-binding sites are compared because the purpose of this approach is to understand the activity of ligands of related targets. MIF (Molecular Interaction Fields) is one measure that can be computed to perform structural comparisons, a structural alignment of all targets is performed and interaction energies are generated by placing probe atoms at each point of the 3D grid the falls in the ligand binding site and placing those energies into a MIF vector. The MIF vector can then be placed into a global matrix where the rows represent targets and columns represent interaction energies at a given 3D grid point. Comparing and clustering the MIF's can be done by analyzing the matrix using a principle component analysis (PCA) (Naumann and Matter, 2002) or by calculating a

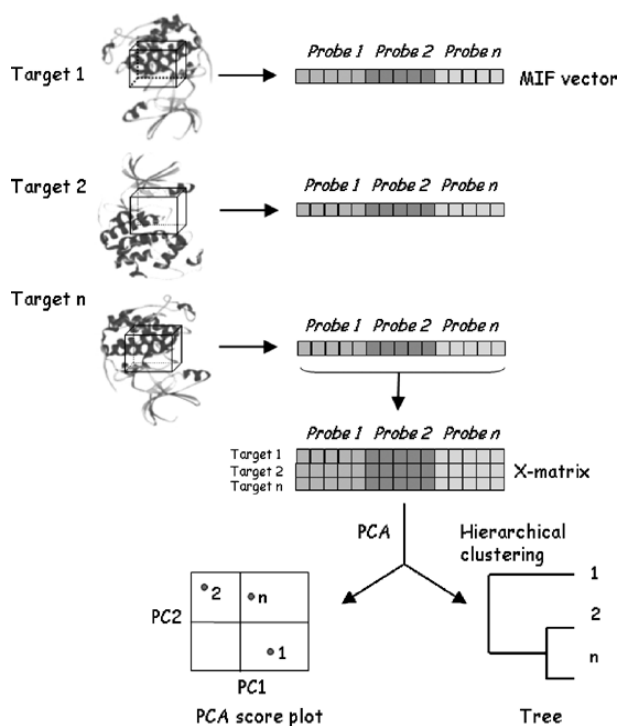


Figure 1.8: Molecular interaction field (MIF)-based clustering of targets (Rognan, 2007).

MIF distance and converting the results into a tree (Hoppe et al., 2006), this is illustrated in Figure 1.8. This type of comparison is very dependent on the structural alignment, the grid resolution and the probe atoms used and cannot be applied to targets in different families but has been successfully applied to protein kinases (Naumann and Matter, 2002) and nuclear hormone receptors (Hoppe et al., 2006) to identify cavity regions that can explain ligand binding and guide the design of compound libraries towards the desirable selectivity pattern.

1.5.3.4 Target-ligand approaches

This approach tries to predict ligands that bind to a particular target by leveraging binding information for other targets without first trying to define a set of receptors (Rognan, 2007). This has been attempted in numerous ways, an example being Bock and Gough (Bock and A, 2005) who combined descriptors of protein-ligand interactions to describe putative ligand-receptor complexes and used machine learning methods to analyze whether a receptor-ligand pair is predicted true or not and Erhan *et al.* (2006) used the same principle but used neural networks and support vector machines to perform their comparisons. They showed

that it was possible to combine a set of receptor descriptors and a set of ligand descriptors into a computational framework that allows for a large degree of flexibility in the choice of descriptors

1.6 Ligand-based approaches

Chemical similarity searching with the goal of target prediction has already been outlined in the chemogenomics approaches section and differs only slightly in focus to find new compounds for a single target, or target to a single ligand. This is performed by comparing a compound's structure to a database of ligands with known targets. The following section will outline a few additional approaches to ligand based approaches of predicting protein ligand interactions. These all still utilize the assumption that similar compounds will be active against the same targets and focus around comparing compounds.

1.6.1 QSAR

Quantitative structure activity relationships (QSAR) are studies performed to understand the quantitative correlation of molecular structure to the binding constant, and thus also predict the properties for novel compounds and help to characterize the spatial features responsible for the changes in activity when comparing drug molecules (Durdagi et al., 2008).

1.6.1.1 Comparative Molecular Field Analysis (CoMFA)

The main focus of CoMFA is to identify a ligand that will have the best affinity to a protein, by analysing a set of compounds that have binding affinity data for a protein target. In short the approach builds statistical and graphical models that relate to the properties of a molecule to its structure. These models are then used to find the activity of novel compounds (Cramer et al., 1988). In CoMFA the molecular structures are first drawn in a rectangular grid that consists of equally spaced lattice points. The electrostatic and steric interaction energies are then calculated by placing a probe atom at each lattice point and an activity model is

produced. The activity model is analyzed and the most important features where steric and electrostatic interactions influence the activity of the compound are selected and represented on a 3-D pharmacophoric map. A problem with CoMFA is that it does not map all types of interactions that take place

1.6.1.2 Comparative Molecular Similarity Indices Analysis (CoMSIA)

This approach is similar to CoMFA but differs only in that it does not map interaction energies using a probe but rather uses distance dependent similarity indices to probe atoms to determine the interaction type and strength. Thus it is able to take all interaction types into consideration.

1.6.2 Structure-based virtual screening

The premise of structure based virtual screening is that the researcher has an already-resolved 3D structure of a protein of interest. In this methodology a “docking program” is then used to dock or fit a computer representation of a small molecule into either the whole or selected area (active site) of the 3D structure of the protein of interest. The program will try to fit this small molecule in all possible orientations, each of these are known as a “pose”. The program will then try to identify the most energetically favorable pose, where each pose is “scored” based on how it complements the structure with regards to shape and electrostatic properties. Poses that indicate a ligand to be a good binder are then given good scores, and the process is repeated on all small molecules that a user is screening against and the results are then ranked based on their scores. The ranked list is then used to select compounds that are predicted to be bio-active against the protein for further study, the process is illustrated in Figure 1.9. If this process manages to be performed with reasonable accuracy it will result in a list of bio-active small molecules that greatly increase the speed of drug development without greatly increasing the costs to the discovery project.

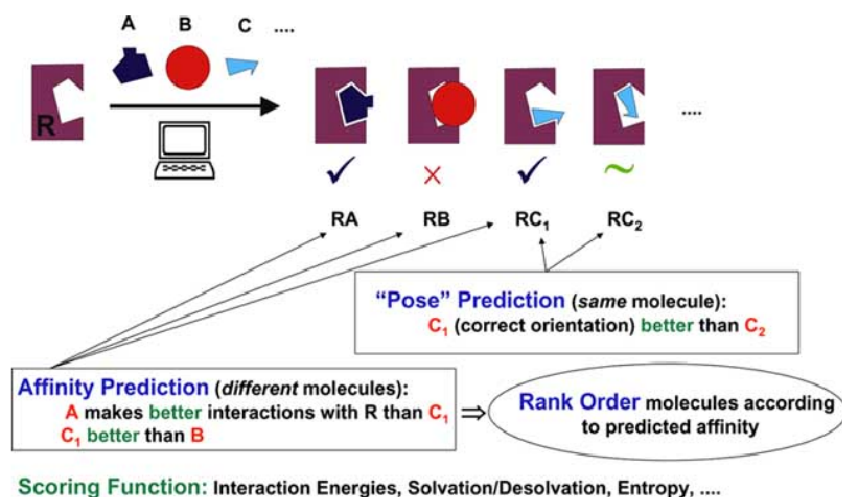


Figure 1.9: Illustration of docking and scoring. R symbolizes a receptor structure, A, B and C represent small molecules to be docked into the receptor (Kroemer, 2007).

1.6.2.1 Basic requirements

In order to perform structure-based virtual screening the receptor structure is needed. Primarily a structure is resolved experimentally using either X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR).

The procedure behind X-ray crystallography Crystallography in brief consists of three steps, the first being the most difficult is to obtain an adequate crystal of the protein in question. The crystal needs to be large, pure in composition and regular in structure with no significant cracks or imperfections. In the second step the crystal is placed in a beam of single wavelength X-rays that produce a regular pattern of reflections while the crystal is slowly rotated, with each orientation the previous reflections disappear and new ones appear and all are recorded. Usually a structure will need multiple data-sets containing tens of thousands of reflections. The last step is to combine this data computationally and add chemical information to create a refined model of the arrangement of atoms in the structure (Smyth and Martin, 2000).

NMR of proteins NMR also requires multiple steps, the first being sample preparation where a large quantity of purified protein product needs to be produced or extracted, the

purified protein is then dissolved in a buffer solution and adjusted to the desired solvent conditions. The second step is called resonance assignment, which is to find out which chemical shift corresponds to which atom, this is done using a technique called sequential walking which combines information from many NMR experiments to validate chemical shift, the procedure varies depending on if the protein has been labeled with carbon-13 and nitrogen-15 or not. The next step is restraint generation, where before a structure can be generated various experimentally determined restraints need to be created, these include restraints on distances and angles. And lastly the structure is generated and validated by the researcher trying to fit as many of the generated restraints as possible (Wuthrich, 1986).

Homology Modelling If the structure of a protein is not available a user can also resort to predicting a 3D structure either through “threading” or homology modeling. The process of threading involves comparing the protein sequence to a database of structures with known folds and building the structure from those results. Homology modeling requires the sequence comparison and similarity to at least one protein that has a 3D structure. Once a structure is derived for a protein it is important to analyze it for potential binding sites which molecules may be bound to.

Pose prediction To find the best pose of a molecule being docked the program usually attempts all the possible orientations and only stores the one that is energetically the best. With the fact that ligands are flexible the program needs to find the best orientation and conformation, this increases significantly the possibilities so the programs usually stop once a certain number of trials have been completed or enough poses have been found with favorable energies. The decision to keep or reject a pose is based on a score that computes the interaction energy of the ligand-receptor. Many programs use a crude “dock score” based on a simple energy function to quickly evaluate poses and then recalculate a “affinity score” using more sophisticated calculations on a smaller set of predictions.

Scoring or affinity prediction For each best pose that was found for a molecule an affinity scoring function is applied to give a binding score of that chemical to the protein. Those scores are then ranked and a list of potential interacting molecules is created. There are two main categories of scoring functions that exist.

Knowledge-guided scoring functions form the first group and these are derived by using statistics of observed inter-atomic contact frequencies and distances in databases of crystal structures of protein-ligand complexes. The assumption is that only the interactions that adhere to the frequencies in the databases favor a positive interaction taking place and increase the overall binding activity, in contrast interactions that have a low frequency in a database are assumed to destabilize the binding thus decrease the overall affinity. The various predictors created include PMF, DrugScore and SmoG, though the major differences are in the size of the training sets and the molecular interactions that are taken into account (Kroemer, 2007).

Energy component methods

form the second major division of affinity scoring methods. These methods are based on the assumption that the total change in free energy of the binding of a ligand to its receptor is made up of the sum of individual contributions.

$$\Delta G_{\text{bind}} = \Delta G_{\text{int}} + \Delta G_{\text{solv}} + \Delta G_{\text{conf}} + \Delta G_{\text{motion}}$$

In the described formula ΔG_{bind} represents the total change in free energy, ΔG_{int} represents the receptor-ligand interactions, ΔG_{solv} represents the interactions occurring between the ligand and receptor with the solvent, ΔG_{conf} represents the changes in conformation of the protein and ligand and ΔG_{motion} represents the change in energy caused by the motions of the ligand and protein during the complex formation. This principle can only truly be applied if you can separate these energies into mutually independent variables which is not really the case with protein-ligand interactions because many of these factors are directly

related and can affect the binding with both positive and negative contributions. The additivity principle has also been proven to not always be applicable in protein-ligand binding (Dill, 1997). Despite the limitations to the accuracy of these methods they are still used because of the quickness of their calculation which is very important when dealing with large databases in high-throughput experiments. This method can also be further divided into two categories those are methods that are based purely on the physical chemistry properties and those that use experimental data to better define the interactions that take place these are called regression-based methods. These assume a linear relationship between the change in free energy and the number of terms (eg. hydrogen bonds, ion pairs, molecular flexibility, contact surface) that characterize the binding. These use crystal structures with binding data to optimize coefficients of the regression equation. Many of the most used scoring functions are developed like this, these include LUDI (Böhm, 1994), ChemScore (Eldridge et al., 1997) and GOLD score (Jones et al., 1995).

1.6.2.2 Challenges to Docking based methods

Despite the large strides taken in this field scientists still face some fundamental challenges to docking and scoring .

Flexibility and Docking Protein flexibility is one of the more difficult problems facing docking algorithms. It has been found in multiple cases that proteins change their conformations when different ligands bind to them (Teague, 2003). This means that docking algorithms that use a rigid structure of a protein will miss the ligands that bind to alternate conformations of the protein. Three main approaches have been used to tackle the problem of protein flexibility, allowing the receptor or parts of the receptor to move during docking, docking a molecule to various conformations of the protein and aggregating the results and prior to docking, averaging the receptor representation. The implementations of these approaches often use various combinations of these approaches to handle flexibility. Despite these approaches the handling of protein flexibility it remains a serious issue that needs to be considered when performing any docking projects.

Water Water molecules often mediate protein-ligand interactions thus failing to include them in the docking calculations will result in the calculated interaction energy of the ligand being too low. Likewise, if a docking program leaves water molecules that are observed in the crystal structure then a ligand that would normally replace the water molecule would not be docked correctly. To appropriately handle water molecules one would need to predict where a water molecule would interact with the protein and ligand and after that determine if the water molecule is actually present at that location.

Tautomers and ionization patterns Tautomers cause a problem in docking because it is not possible to predict which tautomeric state a compound will adopt. Many databases store acids and amines in neutral form, and many of these would be ionized in normal conditions in a cell and therefore would need to be ionized before docking. Although this is not difficult, to compute selecting the correct tautomer is. It is up to the researcher to decide how to deal with this challenge, whether to stick to one single tautomer or generate all possible tautomers for a molecule and dock them all. The value of generating many tautomers is questionable as it could nearly result in a large list of false positives

With all the methods described above it has been shown that there are many ways of using biological information to find chemical ligands that bind to a target protein, to find targets of chemicals known to inhibit an organism and to create and optimize ligands to be effective inhibitors of a target protein. It is not possible to do any performance comparisons of these methods because they do not use the same type of data and each is applied in a different context to answer different scientific questions by providing the answer to the same simple question “Will this compound interact with this protein?”. The choice of approach will always be determined by what question the user is trying to answer and what information is already available. Fortunately this field of science has been in development for many years due to the commercial interest in drug discovery, thus it is possible to perform interaction predictions with a variety of information available even if that information is very limited.

Many of the techniques require proteins and ligands that are already well curated with interaction information to be useful. In recent years large quantities of these reference interactions have become available and thus have increased greatly the value of these approaches, and in years to come as the number of these interactions increases the power of these predictive techniques increases. The largest amount of predictive power will be gained when interaction data regarding less studied protein families becomes publicly available which will unlock the ability to perform predictions on a much greater diversity of proteins thus also revealing new potential targets.

1.7 ChEMBL

ChEMBL is an open access database containing binding, functional and ADMET data for a large number of small drug-like bio-active molecules. The ChEMBL group is based at the European Bioinformatics Institute (EBI) at the Wellcome Trust Genome campus in Hinxton, England led by John.P.Overington and predominantly funded by the Wellcome Trust.

The problem that the ChEMBL group is trying to solve is the difficulty experienced when trying to do research into past drug discovery experiments. The nature of publications is such that chemical structure data is usually published as images making it not possible to search programmatically and proteins are referred to through various synonyms or abbreviations. Additionally most journals do not require the publication of the small-molecule assay results into a public database making the results of the publications only accessible through commercial products.

The core set of activity data that exists in the ChEMBL database was manually extracted from full text peer reviewed publications from various journals including *Journal of medicinal chemistry*, *Bio-organic medicinal chemistry letters* and *Journal of natural products*. *Letters and Journal of Natural Products*. The journals were selected to capture the highest quantity of high quality data at the lowest cost. Above the literature-derived data ChEMBL also contains the structures and annotations of FDA-approved drugs

For each publication the following is included:

- Details of the tested compound, what assays were performed and all target information is abstracted.
- Small molecule structures are drawn in full, in machine readable format including those that are only referred to by name.
- Information regarding the particular salt form being tested is captured.
- The structures are checked for potential problems e.g. unusual valence on atoms, incorrect structures of common compounds.
- The structures are then normalized to a set of rules to create consistency within the database. Compounds are neutralized to get the formal charge to zero where possible, common groups are set to previously decided on representations. Stereochemistry is configured to the naturally occurring configuration unless stated otherwise in the article and common salts are stripped from the compounds and added separately to the database
- All types (including ADMET, functional and binding assays) of assay detail is extracted and activity endpoints values are normalized to improve a users ability to compare values from different assays.
- Protein targets are standardized to be consistent and detailed annotation of targets is handled manually internally by the ChEMBL group.

As of version 16 ChEMBL contains information for 1,295,510 distinct compounds and 9,844 protein targets described in 50,095 publications. There is a data exchange program in place between PubChem BioAssay (Wang et al., 2012) which houses many results primarily of high throughput screening experiments which lack dose-response data (IC₅₀,K_i) as experiments are usually done with a single concentration, but have a significant number of data points so the data between these two databases is distinctly different and complimentary. All ChEMBL assays have been loaded into PubChem and a subset of assays from pubchem

loaded into ChEMBL and clearly marked. Similarly entries have been added reciprocally between ChEMBL and BindingDB (Chen et al., 2001).

ChEMBL uses a web based interface that can be found at <https://www.ebi.ac.uk/chembl/>, this project only deals with the content of chembl so this will not be discussed further.

1.8 Malaria centered bio-activity data sets

There have been various groups in recent years that have released drug screening results against in house chemical databases into the public domain that are generally proprietary data. This is in an effort to stimulate and aid the public sector in the development of either a cure for malaria or discovery of new anti-malarials

1.8.1 GlaxoSmithKline TCAMS data set

In 2010, GlaxoSmithKline(GSK) published the structures of 13533 chemical starting points for antimalarial lead identification that were identified through *in vitro* screening whole cell of 1,986,056 compounds from the GSK screening library on *Plasmodium falciparum* 3D7 strain and subsequently on Dd2.

1.8.1.1 A summary of the methods

The assays (384-well) were prepared with a total concentration of 2 μM of each compound with column 6 as a positive control containing 5 μM of Di-methyl sulfoxide(DMSO) and column 18 with 50 μM artemisinin and μM chloroquine as negative control. The blood cells were added and plates were shaken for 10s to ensure mixing and then incubated at 37C for 72h in an atmosphere of 5% CO₂, 5% O₂, 95% N₂. The screening was completed by evaluating the activity of LDH(Lactate Dehydrogenase) by measuring the level of absorption of reaction mix(143 mM sodium l-lactate, 143 μM 3-acetyl pyridine adenine dinucleotide (APAD), 178.75 μM Nitro Blue tetrazolium chloride (NBT), 286 $\mu\text{g ml}^{-1}$ diaphorase (2.83 U ml⁻¹), 0.7% Tween 20, 100 mM Tris-HCl pH 8.0) after the incubation period.

The decision to incubate for 72 hours was to guarantee that all parasites completed at least one cell cycle and to increase the chance of detecting slow acting and 'delayed death phenotype' inhibitors, some of which were detected, i.e. tetracyclines. Due to the large number of hits many of the concentration curves were estimated using the LDH assay instead of the standard hypoxanthine incorporation assay which is regarded as the standard for anti-malarial concentration curves. These were demarcated as XC_{50} instead of the usual IC_{50} . To measure cytotoxicity the hits were screened at 5 times the screening concentration against human hepatoma HepG2 cells (Gamo et al., 2010).

1.8.2 Novartis-GNF Malaria Box

The Novartis-GNF Malaria box is a selection of compounds from GNF's non-proprietary chemical libraries that were screened for proliferation inhibition activity of *P.falciparum* strain 3D7 in human erythrocytes. The data set contains the structures and screening results of 5600 compounds, that were tested in dose response and confirmed to inhibit *P.falciparum* growth by at least 50% at the highest screened concentration being 12.5 μM . Activity was also measured against the multi-drug resistant W2 strain and in addition to this a human cell cytotoxicity screen was completed using the Huh7 human hepatocellular carcinoma cell line to give indication as to the 'promiscuity' of the hits.

The data sets were created by testing *P. falciparum* strains 3D7 and W2 in an erythrocyte-based infection assay for susceptibility to inhibition of proliferation using the malaria box compounds mentioned above. Compounds were screened in a 12 point dose-response (1/2 log serial dilutions) assay in 1536-well format and concentrations ranged from either 12.5 μM to 0.0001 μM . Parasite cultures (8 μL) at 0.3% parasitemia and 2.5% hematocrit were treated with a compound for 72 hours under low oxygen conditions. After 72hr parasite growth is determined by measuring the nucleic acid content of the parasites with the fluorescent dye SybrGreen and plates were read on an Envision plate reader (Perkin Elmer). Current antimalarials were used as reference compounds.

Compounds were screened against the Huh7 human hepatocellular carcinoma cell line in a

12 point dose-response (1/2 log serial dilutions) in 1536 well format at concentrations ranging from 100 μM to 0.0003 μM . Cells are seeded at 500 cells per 5 μL media per well. Compounds were transferred the next day and cells are cultured for 72 hr to match the incubation length during the *P. falciparum* proliferation inhibition assay. Cellular viability was assessed using Cell Titer Glo (Promega). (K Gagaring)

1.8.3 St Judes Children’s Research Hospital Malaria data set

This data set that was released by the St Judes Children’s Research Hospital contains the details of the effectiveness of almost 310,000 chemicals of which 1,100 are new compounds with confirmed activity against the malaria parasite, 172 were studied in detail which lead to the identification of more than a dozen families of possible candidates. In this study, investigators surveyed the hospital’s library of compounds looking for those effective against the entire malaria parasite. Scientists tested the chemicals against the *Plasmodium*. The work led St. Jude researchers to three families of molecules, including two believed to act against new targets. Investigators hope to have a new drug in the clinic within a decade.

1.9 Discovery

Discovery is web-based system developed to be a resource for researchers to be able to mine information on malaria proteins and predicted ligands, as well as perform comparisons to the human and mosquito host characteristics. Protein features used include: domains, motifs, EC numbers, GO terms, orthologs, protein-protein interactions, protein-ligand interactions and host-pathogen interactions among others (Joubert et al., 2009). The Discovery system has gone through a major update to improve its functionality, it was rewritten in Java with a focus on making it easily update-able. The reason being that the data sources Discovery uses to make its comparisons are constantly being expanded and new entries are consistently being added or changed, these updates are too frequent to be handled manually thus an automated solution was required. The importance of keeping the data as current as possible

is that new data may allow for new associations to be made with malaria proteins which could lead to new drug targets or leads being identified.

1.10 Problem Statement

The current dependence on artemisinin based combination therapies and the emergence of artemisinin resistance in Asia creates the need to the continuation of the search for alternative drug targets and lead compounds to treat malaria. However all drug development projects require a large amount of capital input, both financially and man hours, making failed drug development projects very costly as failure causes all capital invested to be lost. A great amount of value can be realized by reducing the number of failures by improving selection of drug targets or leads prior to committing resources. This can be achieved by aiding drug target selection with *in silico* techniques.

With the increasing quantities of bio-active molecule-data available to the public an opportunity exists to leverage this data to improve the selective ability of *in silico* target selection. By adding possible protein-ligand interactions to the list of post-genomic data available you gain possible insight into a proteins activity and should a target be selected it may provide a list of chemical compounds that can be used as start points for screening efforts. The problem that needs to be addressed is how to integrate this type of data with other malaria data to make this leveraging possible.

1.11 Aims

The aim of this research project was:

- To expand on the protein-ligand interaction prediction ability in Discovery 2.0.
- To find and incorporate alternative bio-active molecule databases.
- To improve accessibility to malaria data using ligand-based approaches.
- To improve on the platform to do chemical based searches against malaria proteins.

- Aid in the improved characterization of malaria proteins by predicting their chemical inter-actors.
- Improve on the Discovery platform used for searching malaria data.

Chapter 2

Methods

2.1 Overview

The goal of this project has been to expand the Discovery systems ability to identify putative protein-ligand interactions for the malaria parasite *Plasmodium* proteins, previously Discovery was designed to perform a protein BLAST of *Plasmodium* proteins against Drugbank database that houses the protein target information of FDA approved drugs. Although this is very useful data a limitation is that it is static and has not been updated since Drugbank version 3.0 which contains 4,229 protein targets and are of FDA approved drugs only. With the focus of the new version of Discovery being for it to be as up to date as possible it was decided to search for and add additional bioactive sources to use for comparisons.

2.2 ChEMBL database integration

The ChEMBL database is available online in various formats. All compounds are available as an .sdf file and all protein targets can be downloaded as a .fasta file. Alternatively the entire database can be downloaded in Oracle, MySQL and PostgreSQL versions from “<https://www.ebi.ac.uk/chembl/downloads>”.

For integration into Discovery 2.0 which uses a MySQL database, the MySQL version

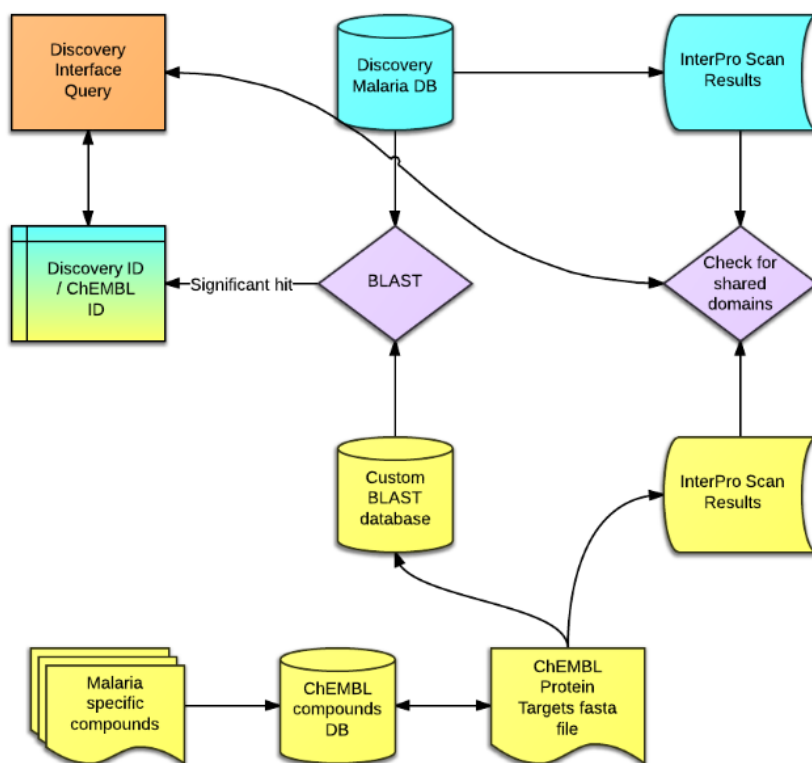


Figure 2.1: Flow Diagram of Protein-ligand interaction prediction in Discovery 2.0. The different sources are represented with different colours, ChEMBL and its components in yellow and Discovery in teal with merged data represented by the gradient

of ChEMBL was downloaded along with the .sdf and .fasta files. The first aspect that was addressed was how Discovery connect the chemical data to malaria proteins, this was done using BLAST and domain matching which is discussed later. To prepare these associations the ChEMBL entries were compiled into a single reference table that contained the ChEMBL id of the target protein in one column and its associated compounds ChEMBL id in the second column. There is a many-to-many relationship between the chembl protein ids and the ChEMBL chemical ids as one protein can have many binding ligands and a single compound can possibly interact with many different proteins, thus the need to create a simple indexed table storing the associations to increase the speed relevant queries.

2.2.1 Chemical Data Integration

Chembl contains 1,324,941 distinct chemical compounds with detailed information for each record. There was limited value in adding all of this information into Discovery and a focus was maintained on keeping the display simple and allow external links to the ChEMBL website through the results pages, so it was decided that apart from the basic information just the structural data of each compound and the association data is stored in Discovery. The chemical searching in Discovery is handled by a third party tool called JQuery by ChemAxon, thus all the chemical data was imported into the JQuery plug-in for it to create its own indexed database for quick structure based queries.

This integration allows then the use of the Marvin sketch tool to draw compounds in the Discovery web page query the entire chemical database.

2.2.2 Protein Data Integration

ChEMBL has 9,844 protein targets available. The protein data was stored in the .fasta file and a simple table containing the proteins sequence was created in the Discovery database for rapid access when performing the sequence alignments in the proteins page.

2.2.3 Assay data integration

Each ChEMBL entry of every chemical or protein entry has an ChEMBL assay entry associated via a foreign key this is a reference to a document where in the interactions are described. This data was also incorporated into Discovery where by the short description is available and external links exist to take the user to the entry on the ChEMBL website.

2.3 Protein Matching using BLAST

BLAST is the tool used to measure the similarity between the proteins existing in the ChEMBL targets database and the malaria related proteins.

2.3.1 BLAST

The Basic Local Alignment Search Tool (BLAST) was developed to perform rapid sequence comparisons to identify regions of similarity in sequences and is one of the most widely used bio-informatics tool available. BLAST approximates alignments that are optimized by a measure of local similarity, the maximal segment pair (MSP) score. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. (Altschul et al., 1997)

2.3.2 The BLAST Algorithm

The BLAST algorithm can be summarized as follows

- Remove low-complexity region or sequence repeats in the query sequence.
- Create a k-letter word list of the query sequence.
- List the possible matching words.

- Organize the remaining high-scoring words into an efficient search tree.
- Repeat step 3 to 4 for each k-letter word in the query sequence.
- Scan the database sequences for exact matches with the remaining high-scoring words.
- Extend the exact matches to high-scoring segment pair (HSP).
- List all of the HSPs in the database whose score is high enough to be considered.
- Evaluate the significance of the HSP score.
- Make two or more HSP regions into a longer alignment.
- Show the gapped Smith-Waterman local alignments of the query and each of the matched database sequences.
- Report every match whose expect score is lower than a threshold parameter E.

2.3.3 Application

In the context of Discovery and specifically this project there was no interest in using BLAST on the standard BLASTp database but rather to only use the chembl targets database. To do this the following steps were taken:

1. The latest version of blastp was downloaded from “[ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/'](ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/) .
2. All the protein sequences from the ChEMBL targets database and the malaria proteins were converted into two fasta files.
3. Using Formatdb a custom database was created of the ChEMBL fasta file.
4. The malaria targets were then BLASTed against the ChEMBL database and a resulting BLAST file was obtained containing all the BLAST hits of the respective proteins

5. The relevant information was then parsed into a MySQL database to store the resulting matches as well as their E-values to make it usable by the Discovery system.
6. The file is also setup to re-run when a new version of ChEMBL is released, it will then automatically run this procedure again and generate updated results.

2.4 Protein Matching by Domain

The concept that similar proteins are likely to have similar function is further emphasized by the modular nature of protein sequences, small sequence fragments often occur throughout a family of proteins and can be categorized as functional/structural or binding domains. These domains carry the functional part of the protein, proteins may be made up of numerous domains to allow it to exhibit a specific task. If this reasoning is true then there lies the potential that proteins can be quite distant in sequence similarity but if they carry the same functional domains they are still likely to share the same activity. This feature has aided largely in identifying protein-protein interactions and is used here in a similar way to find proteins sharing ligand interaction properties. The concept is defined by Jadwin *et al.* (Jadwin et al., 2012) “domainomics” to draw attention to the potential of using domains and their motifs as tools in proteomics. They propose that the accumulation of domain–motif binding data could ultimately provide the foundation for domain-specific interactomes, which will likely reveal the underlying substructure of protein networks as well as the selectivity and plasticity of signal transduction (Jadwin et al., 2012).

2.4.1 InterPro

There has been much development in the area of protein domain identification and several signature recognition methods have evolved to address different sequence analysis problems, resulting in rather different mostly independent databases. Diagnostically, these resources have different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods. Thus, for best results, search strategies should ideally

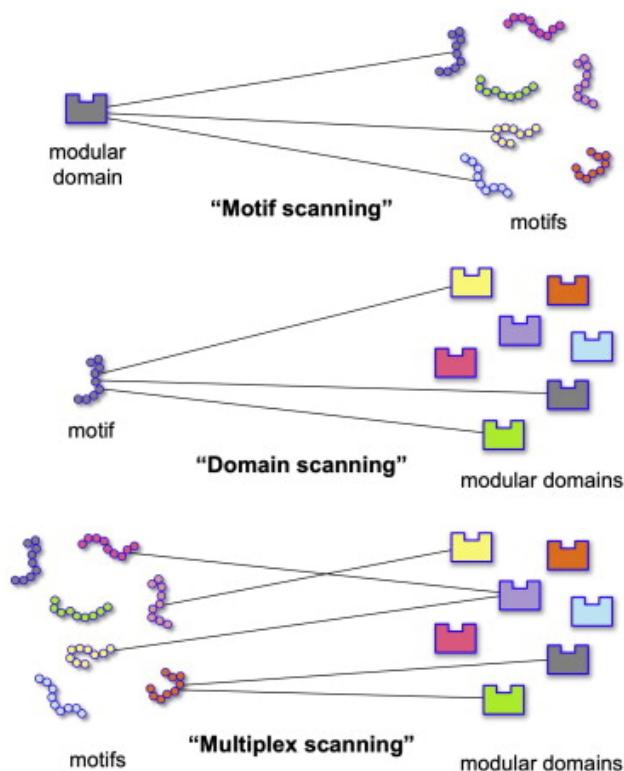


Figure 2.2: This figure illustrates the three basic assay designs for studying interactions between modular protein domains and short peptide motifs. Top: in motif scanning, a domain of interest is used to probe a library of peptide motifs or proteins containing binding motifs, typically to define domain specificity or identify possible binding proteins. For example, an immobilized domain can be used as bait in pull-downs. Middle: in domain scanning, a motif of interest is used as a probe to screen a set of domains or domain-containing proteins. Bottom: multiplex scanning simultaneously assesses interactions between many ligands and domains, providing the specificities of domains within a domain–motif interaction map. Multiplex scanning can be designed as an expanded version of domain or motif scanning, or as a “library to library pull-down” to screen for binding modules (Jadwin et al., 2012).

combine all of them. This is the area which the InterPro group addressed, their mandate was to create a layer to allow users to access all relevant domain databases and methods of sequence analysis in one single interface (Hunter et al., 2012).

InterPro categorizes each entry into one of a number of types which tell you what you can infer when a protein matches the entry.

- **Family:** A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions, similarities in sequence, or similar primary, secondary or tertiary structure. A match to an InterPro entry of this type indicates membership of a protein family.
- **Domain:** Domains are distinct functional, structural or sequence units that may exist in a variety of biological contexts. A match to an InterPro entry of this type indicates the presence of a domain.
- **Repeat:** A match to an InterPro entry of this type identifies a short sequence that is typically repeated within a protein.
- **Site:** A match to an InterPro entry of this type indicates a short sequence that contains one or more conserved residues. The type of sites covered by InterPro are active sites, binding sites, post-translational modification sites and conserved sites.

The InterPro database integrates PROSITE (Bucher et al., 1996), PRINTS (Scordis et al., 1999), Pfam (Sonnhammer et al., 1998), ProDom (Corpet et al., 1998), SMART (Schultz et al., 1998), TIGRFAMs (Haft et al., 2001), PIR superfamily, SUPERFAMILY (Eddy, 1998) Gene3D (Buchan et al., 2002), PANTHER (Mi et al., 2010) and HAMAP (Bucher et al., 1996) databases. As previously discussed there are differences in the manner in which these databases need to identify their respective domains. A brief description of the various identification methods each database uses for domain matching

PROSITE Patterns: Some amino acid patterns can be formed into regular expressions, this applies to highly conserved sequences.

PROSITE Profiles: There are a number of protein families as well as functional or structural domains that cannot be detected using patterns due to their extreme sequence divergence, so the use of techniques based on weight matrices (also known as profiles) allows the detection of such proteins or domains. A profile is a table of position-specific amino acid weights and gap costs.

HAMAP profiles: HAMAP profiles function in a similar way to PROSITE profiles but are used specifically to identify protein families from Bacteria and Archaea and propagate annotation to them.

PRINTS: The PRINTS database houses a collection of protein family fingerprints. These are groups of motifs that together are diagnostically more powerful than single motifs by making use of the biological context inherent in a multiple-motif method. The fingerprinting method arose from the need for a reliable technique for detecting members of large, highly divergent protein super-families.

PFAM: Pfam contains curated multiple sequence alignments for each family and corresponding Hidden Markov Models (HMMs). Profile Hidden Markov Models are statistical models of the primary structure consensus of a sequence family. The construction and use of Pfam is tightly tied to the HMMER software package (Eddy, 1998).

PRODOM: ProDom is a database of protein domain families obtained by automated analysis of the SWISS-PROT and TrEMBL protein sequences. It is useful for analysing the domain arrangements of complex protein families and the homology relationships in modular proteins. ProDom families are built by an automated process based on a recursive use of PSI-BLAST homology searches.

SMART: SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. These domains are extensively annotated with respect to phylogenetic distributions, functional class, tertiary structures and functionally important residues. SMART alignments are optimized manually and following construction of corresponding Hidden Markov Models (HMMs).

TIGRFAMS: TIGRFAMS are a collection of protein families featuring curated multiple

sequence alignments, HMMs and associated information designed to support the automated functional identification of proteins by sequence homology.

PIR SuperFamily: PIR SuperFamily (PIRSF) is a classification system based on evolutionary relationship of whole proteins using HMMs.

SUPERFAMILY: SUPERFAMILY is a library of profile Hidden Markov Models that represent all proteins of known structure, based on SCOP.

GENE3D: Gene3D is supplementary to the CATH database. This protein sequence database contains proteins from complete genomes which have been clustered into protein families and annotated with CATH domains, Pfam domains and functional information from KEGG, GO, COG, Affymetrix and STRINGS.

PANTHER: The PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System was designed to classify proteins (and their genes) in order to facilitate high-throughput analysis also using HMMs.

2.4.2 InterProScan

InterProScan is a tool that combines the different protein signature recognition methods described above into one resource (Quevillon et al., 2005). The September 2012 release has 23,792 entries which are comprised of 15,865 Families, 6,834 Domains, 269 Repeats and 824 Sites. InterProScan takes a protein sequence and returns a list of domains identified from the various databases.

An overview of sequence of events once a sequence is submitted is described below. As outlined in the InterProScan documentation.

1. The sequence(s) is checked for illegal characters and reformatted if necessary. If it is a DNA sequence, it is translated into 6 frames, according to how the user has configured it. Based on the users input, the sequence file may be split into smaller files.
2. A CRC64 checksum is then calculated for each sequence. A checksum allows the program to check whether that sequence is already present in the InterPro database. If it

is, the models which match to it will already have been calculated and so, to save time returns these results to the user, rather than having to re-run all the searches.

3. If the checksum finds a match in the XML file, InterProScan stores the details of the match. If there is no match, the novel sequence is put into a file with the extension ".nocrc" ready for searching the model databases.
4. InterProScan will then launch whatever applications have been specified by the user (e.g. a HMMer search against TIGRFams). The command-line provided for each database search can be found in the corresponding .conf file. This step will produce a raw output (".output") file in the working directory. Some databases (such as PANTHER and BlastProDom) also produce temporary files in the process but these can be disregarded.
5. Once the initial model search has completed, InterProScan will read in the raw output and apply post-processing to filter out incorrect hits. This filtering steps varies from database to database (e.g. there is none in TigrFAMs but a lot in Pfam).
6. After post-processing, the results for each chunk will be output into a file called merged.raw together with the precomputed match information from the checksum sequences.
7. All the raw files are merged together into a single merged.raw file in the top-most level of the working directory for the run. This is then converted (if necessary) into whatever format the user requested.

2.4.3 Application

The procedure to apply this sort of comparison in Discovery was to create a FASTA file containing the sequence information into each of protein targets existing in the ChEMBL database and the collection of *Plasmodium* proteins that were already in the Discovery system and run them through the InterProScan program and create domain profiles for each protein

and store them in a SQL database. A SQL query was then written to search through the database using the domains of the query protein and return matching ChEMBL proteins that share domains with the *Plasmodium* proteins.

2.5 Clinical Trials exploration

Given that the main objective of this project was to create a tool to aid in the discovery of protein-ligand interactions there was some value in exploring the protein-ligand interactions that are already proven to exist and are exploited as therapeutics. The purpose of adding this type of data was to provide users with a summary of up to date clinical trials data pertaining to malaria research but also to act as a starting point for ligand based searches. The clinical trials data is gathered from clinicaltrials.gov website.

2.5.1 Clinical Trial

In the development of new drugs there are various phases the drug goes through before it becomes available to public for consumption. Drugs first need to be identified in what is known as the pre-clinical stage where *in vitro* and *in vivo* experiments are carried out on model organisms to identify potential candidates that show low toxicity and high efficacy in treating a particular disease. Subsequently to selection the drug goes through clinical trial that can be broken down into various phases of the drug development process.

- Phase 0 : A small study usually on 10 patients to test pharmacodynamics and pharmacokinetics of drug using subtherapeutic dosages.
- Phase 1: A study of between 20-100 patients with the aim to test if a drug is safe for efficacy testing and to identify safe dosages.
- Phase 2: A study on 100-300 healthy patients with the aim to determine the efficacy of the drug.
- Phase 3: A study on 300-3000 patients to determine the drug's therapeutic effect.

- Phase 4: At this stage the drug is available on the market and this represents monitoring of possible long term effects of the drug

2.5.2 Clinicaltrials.gov

ClinicalTrials.gov is a web-resources that houses information on publicly and privately supported research covering a wide range of diseases and treatments. The web site is maintained by the National Library of Medicine(NLM) at the National Institutes of Health (NIH). Information is supplied and updated by the principal investigator or sponsor of the clinical study and in general studies are submitted when they begin and information is updated throughout the study.

Clinicaltrials.gov was created as a result of the Food and Drug Administration Modernization Act of 1997 (FDAMA) in the USA. FDAMA required the U.S. Department of Health and Human Services to establish a registry of clinical trials information for both federally and privately funded trials conducted under investigation new drug applications to test the effectiveness of experimental drugs for serious or life-threatening diseases or conditions. NIH and the Food and Drug Administration (FDA) worked together to develop the site, which was made available to the public in February 2000. For drugs to be cleared by the FDA they need to comply with many regulations and one currently being that they register their clinical trials on this web resource, thus making it a good resource for this type of data housing over 130,000 clinical trials taking place in more than 180 countries. This is, however, not an exhaustive list as certain drugs are not compelled to register their trials and companies not aiming for FDA approval also have no need to register their trials.

2.6 Discussion

The methods as described above were utilized to address the requirements of detecting protein-ligand interactions in previously uncharacteristic proteins. The focus was to generate links between proteins that have ligand binding information with malaria proteins that

have none. By using BLAST and InterProScan which are both tried and tested scientific methods there is little learning that a researcher will need to complete before being able to use this tool effectively. This will hopefully increase in the usability of the system as there are no novel techniques or algorithms that need to be understood in order to interpret the results that are generated.

The utilization of the clinical trials data provides an insightful look into drug research for malaria. In context of protein-ligand interactions, it provides a list of ligand start points for searching the chemical space of malaria by beginning with compounds that are very likely to be interacting with malaria proteins, thus allowing users with no data to test against a small set to work with.

Chapter 3

Results

The following section will illustrate the improvements made to the Discovery systems protein-ligand interaction prediction ability as well as the changes to the interface. A brief summary of the old system will be explained to illustrate the changes and improvements made. Subsequently a summary of the current data statistics will be provided to illustrate the expansion to the knowledge base used to identifying protein-ligand interactions where little or no previous information existed.

3.1 Discovery v1.0

The focus of this summary will be on the protein-ligand interaction searching capability of the old system, so that a direct comparison can be made between the old and new systems. To make it easier to compare the same proteins and compounds have been run through both systems which will appear on each of the screen shots taken from the systems. The protein chosen for protein searching was lactate-dehydrogenase (LDH) which is one of the proteins used as a case study. LDH is an enzyme responsible for catalysis of the interconversion of pyruvate and lactate, and a *Plasmodium* specific LDH exists and has been well characterized (Makler et al., 1993). For the ligand searching spermidine was chosen, it is a molecule that is found in ribosomes and has a strong relationship with cell survival and binds and precipitates DNA (Wan and Wilkins, 1993). Both of these exists in most organisms and have been well

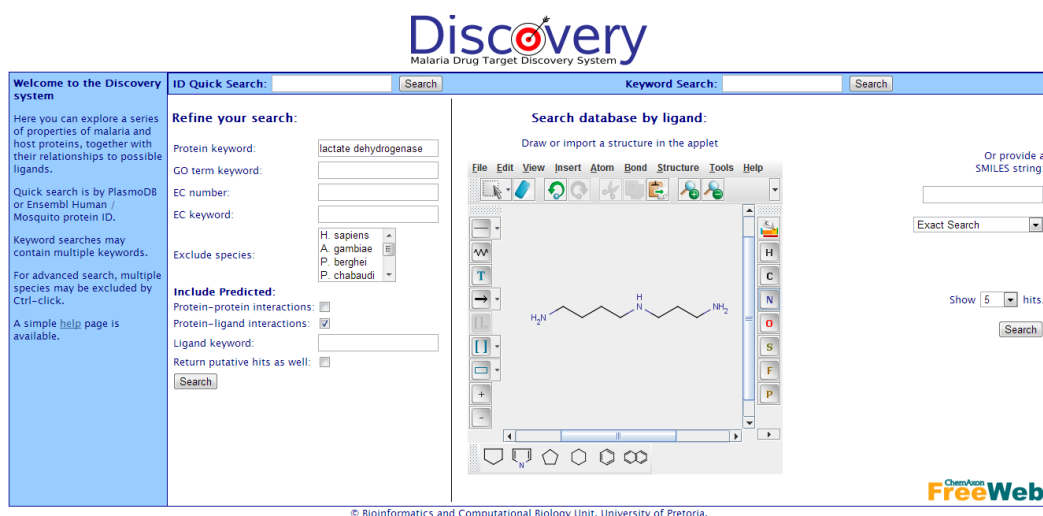


Figure 3.1: A screen capture of the landing page of Discovery 1.0 with the compound spermidine sketched with the Marvin sketch tool.

studied thus definitely having matches in both systems. Please note spermidine does not show any evidence of having binding affinity with LDH.

3.1.1 Primary search page:

Searching for interactions of a protein could be done by entering the protein descriptor of your choice and selecting the include predicted protein-ligand interactions check box then performing the search. A snapshot is provided in Figure 3.1 showing the layout and selection options.

Alternatively if a user was interested in finding potential protein matches to his ligand he has an option to draw the molecule in the Marvin sketch tool developed by ChemAxon, or he could provide a SMILES string for his compound or provide the compound name.

3.1.2 Results of search by protein name

When searching by protein name a list of matching proteins is provided the user to choose from and once selected an entry from the list, the protein information page appears in Figure 3.2. The protein information page provides detailed information of a broad range of areas. Specifically the ligand interactions section displays a hierarchy list that allows you to view



Discovery
Malaria Drug Target Discovery System

Summary | Orthology | Function | KEGG Metabolism Maps | Structure | Protein interactions | Ligand interactions | Host Pathogen interactions

This page shows possible protein-ligand interactions for the *P. falciparum* protein PF13_0141. The predicted protein-ligand interactions are based on KEGG annotations of protein PF13_0141, and on BLAST searches against the PDB and DrugBank databases. Currently, prediction is based purely on BLAST results, and no similarity or other scoring is implemented. This will be improved in future releases.

Ligand interactions | Ligand interactions of homologs

- By ligand
- By source
 - KEGG
 - DrugBank BLAST
 - DBSEQ 2161: L-lactate dehydrogenase: E-value 0.0
 - Nicotinamide-Adenine-Dinucleotide
 - 3-Hydroxyisoxazole-4-Carboxylic Acid
 - 4-Hydroxy-1,2,5-Oxadiazole-3-Carboxylic Acid
 - Oxalate Ion
 - 4-Hydroxy-1,2,5-Thiadiazole-3-Carboxylic Acid
 - Oxamic Acid
 - Naphthalene-2,6-disulfonic acid
 - 3,7-DIHYDROXYNAPHTHALENE-2-CARBOXYLIC ACID
 - DBSEQ 2184: L-lactate dehydrogenase: E-value 2e-91
 - DBSEQ 2211: Lactate dehydrogenase: E-value 1e-75
 - DBSEQ 2212: Malate dehydrogenase: E-value 4e-61
 - DBSEQ 1717: Malate dehydrogenase: E-value 3e-58
 - DBSEQ 2269: Malate dehydrogenase: E-value 3e-58
 - DBSEQ 2229: L-lactate dehydrogenase: E-value 4e-45
 - DBSEQ 2160: L-lactate dehydrogenase: E-value 5e-41
 - DBSEQ 2234: L-lactate dehydrogenase: E-value 1e-40
 - DBSEQ 748: L-lactate dehydrogenase C chain: E-value 8e-40

Figure 3.2: The results page of a protein view with the Ligands section selected, expanding specifically on the top Drugbank hit to indicate its matching compounds

the compounds either based on their sources or by listing the ligands and making the relevant protein detail available by clicking the ligand name you are interested in. In previous version of Discovery contains the sources KEGG, DrugBank blast, PDB Blast, SMID and MSD, none of which appear in the new Discovery system because of the switch to the ChEMBL database as a single source of bio-active molecule information. When drilling down through the list certain of the results are linked to the sources website but vast majority of the results just show the ligand name. When drilling down through the “By ligand” list the final step is information regarding it’s target protein or domain which is linked to the sources web page. Almost all protein entries here have active links.

3.1.3 Results of search by ligand

When searching by ligand a user has to draw the molecule he is interested in, enter its SMILE string or enter it in by name, then depending on what type of search was selected ie, similarity, exact, fragment search etc, and how many results the user wants displayed. A results page will appear as shown in Figure 3.3 indicating all the best matches in descending order, in the spermidine case there were three matches one for Drugbank, one for KEGG and one for PDBligands.

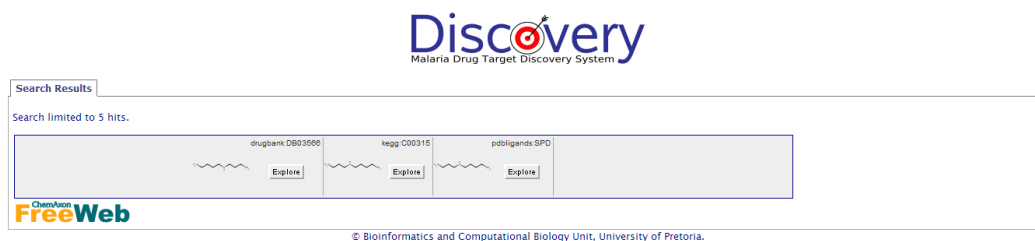


Figure 3.3: The results of a ligand based search in Discovery 1.0.



Figure 3.4: The detail page after finding a molecule of interest within Discovery 1.0.

When a user selects one of the compounds of interest, he is directed to the information page of that ligand as illustrated in Figure 3.4. All relevant ligand information is provided for the user to examine, and a tab appears labeled “Possible Protein Interactions” which is based on a BLAST run against all malaria proteins using the respective protein sequence found in the selected database that has been indicated to interact with spermidine as the query. It is important to note here that the Drugbank and KEGG entries for spermidine had no possible protein interactions and the PDBligand entry had 101 possible matches, thus showing one of the limitations of using multiple bio-active molecule databases. In this case it shows that entries exist for each that contain different data which could possibly cause a user to miss potential interactions.

This was the extent of the first version of Discovery's ligand interaction section. Improvements made to the system will be discussed in the following section.

3.2 Discovery v2.0

The following sections will describe Discovery 2.0's protein-ligand searching capabilities with some comparisons drawn to the previous version. It is important to restate that the objectives of this project were to increase the extent in which the system could identify protein-ligand interactions through increasing available data sources and to retain the ease of use of the platform for the user to access the biochemical data stored in Discovery.

3.2.1 Chemical Searching

In very much the same way as Discovery 1.0 there are primary approaches the user can take in using Discovery to identify protein-ligand interactions, starting with a compound of interest or a protein of interest. The first described is a ligand based approach which is appropriate if a user has a chemical compound of interest and would like to potentially identify a protein partner that it interacts with. On the Discovery home page there are the different search criteria available, a basic search (protein-based search) function allowing a user to search the database by entering either the Protein ID, the Uniprot accession number or by giving a protein alias or name. The second option is the advanced search tab where a user is able to search the entire database by specifying filtering criteria which then returns a subset that meets all the criteria he requested (discussed later). The third search option available is the chemical search page, the user can use any of 3 different search criteria, either draw the 2D chemical structure, enter in the SMILES string or otherwise enter in the compound name to perform a text based search of the compound name. The applet used for the sketching and searches is MarvinSketch by ChemAxon. The Marvin Applet handles many different search approaches, a comparison of available search types is visible in Figure 3.5. By definition, the examined molecule is called a target, the structure we are looking for is called a query, and

Search type	Search feature							
	Similarity	Tests if target contains query	Tests if query contains target	Full fragment coverage	Exact topology matching	Exact stereo matching	Exact atom features matching	Exact bond matching
SUBSTRUCTURE	n/a	✓	✗	✗	✗	✗	✗	✗
SUPERSTRUCTURE	n/a	✗	✓	✗	✗	✗	✗	✗
FULL_FRAGMENT	n/a	✓	✗	✓	✗	✗	✗	✗
FULL	n/a	✓	✓	✓	✓	✗	✗	✗
DUPLICATE	n/a	✓	✓	✓	✓	✓	✓	✓
SIMILARITY	✓	n/a	n/a	n/a	n/a	n/a	n/a	n/a
MARKUSH_MCS	n/a	For the MCS atoms	✗	✗	✗	✗	✗	✗

Figure 3.5: A comparison of available search types available through the ChemAxon search tool, this shows the contrasting properties of each of the search types available

a target molecule matching the query structure is called a hit. There are 5 subcategories of searching available in the jQuery applet:

- Substructure Search : The most common search type performed and searches for whether one molecular structure contains the other one as a substructure or not .
- Exact/Full Search : A full structure search finds molecules that are equal (in size) to the query structure. (no additional fragments or heavy atoms are allowed.) Molecular features (by default) are evaluated the same way as described above for substructure search.
- Exact/Full Fragment Search : search is between substructure and full search: and the query must fully match to a fragment of the target. Other fragments may be present in the target, they are ignored. This search type is useful to perform a "Full search" that ignores salts or solvents beside the main structure in the target.
- Similarity Search : similarity concept is based on hashed binary chemical fingerprints with Tanimoto metrics and is only possible on database searches (Discovery was prepared in such a way to be accessible through this search method).
- SuperStructure Search : search is the opposite of substructure search: It searches for those target molecules which can be found in the given superstructure query (in this case

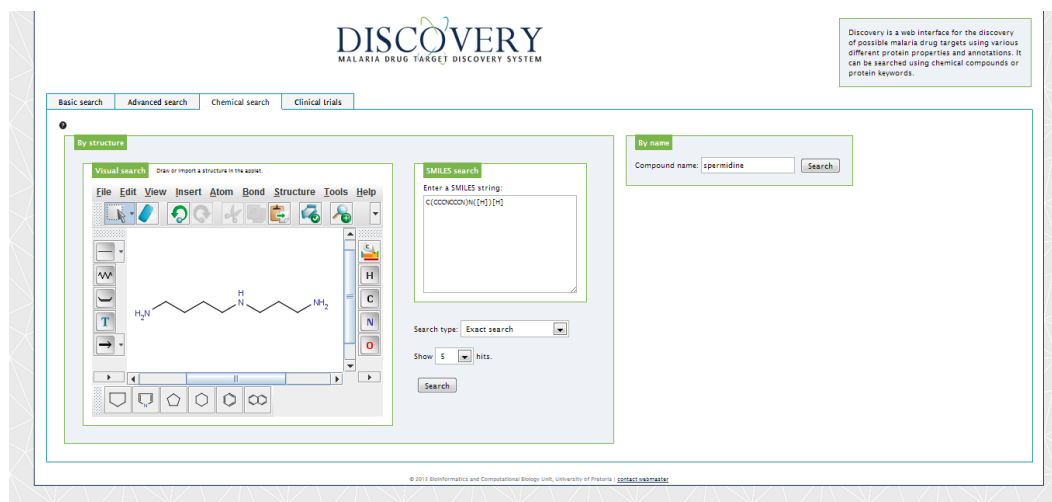
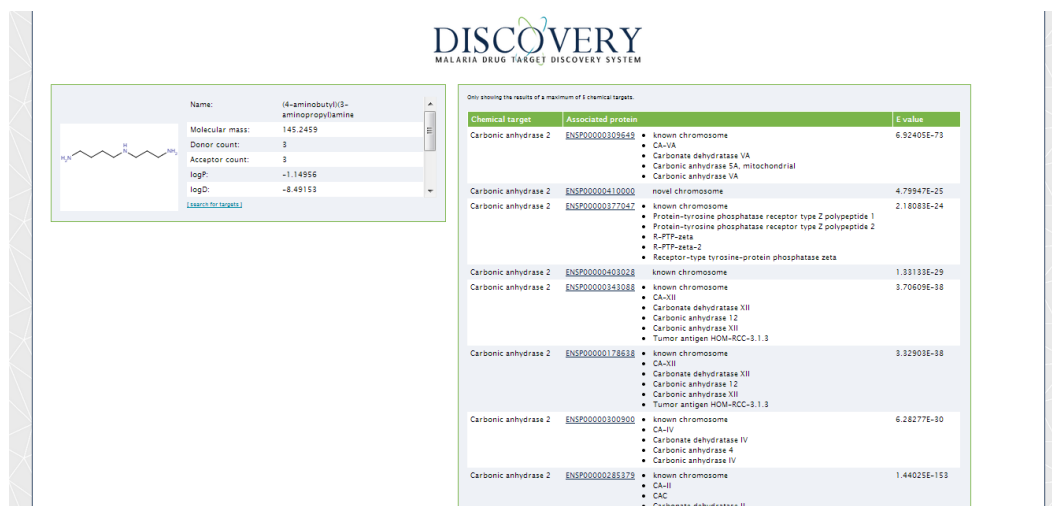


Figure 3.6: The Chemical Search Page: Spermidine has been drawn, the SMILES string and compound name entered, only one of these is necessary to perform a search.

the roles of the query and target molecules are simply exchanged, so query properties should be specified to the target).

Upon submitting a chemical compound to query, the resulting page (Figure 3.7) is a list of compounds from the ChEMBL database matching the search criteria specified. The resulting list contains various information about the chemicals found to match, including: the drawn stereotypical structure, the compound name, molecular mass, donor count, acceptor count, Log P, Log D, Ring count, rotatable bond count, and boolean values for Lipinski's rule of 5, and lead-likeness. The information is intended to allow a researcher enough information to deduce if the compound is similar enough for his liking. Each result also contains a link labeled "[search for targets]" that when clicked enables a new block of information that lists all ChEMBL targets that have interactions with the selected compound. Each of these proteins have then links to the Discovery entry for that protein which will be described in detail in the next section. The goal is thus to potentially aid in identifying a possible target protein of a query compound.



DISCOVERY
MALARIA DRUG TARGET DISCOVERY SYSTEM

Name: (4-aminobutyl)3-aminopyridinium
 Molecular mass: 145.2459
 Donor count: 3
 Acceptor count: 3
 logP: -1.14956
 logD: -8.49153
[\[Search for targets\]](#)

Only showing the results of a maximum of 2 chemical targets.

Chemical target	Associated protein	E value
Carbonic anhydrase 2	ENSP00000105649 <ul style="list-style-type: none"> known chromosome CA-VA Carbonate dehydratase VA Carbonic anhydrase VA, mitochondrial Carbonic anhydrase VA 	6.92405E-73
Carbonic anhydrase 2	ENSP00000410000 <ul style="list-style-type: none"> novel chromosome 	4.79947E-25
Carbonic anhydrase 2	ENSP00000377047 <ul style="list-style-type: none"> known chromosome Protein-tyrosine phosphatase receptor type 2 polypeptide 1 Protein-tyrosine phosphatase receptor type 2 polypeptide 2 R-PTP-zeta R-PTP-zeta-2 Receptor-type tyrosine-protein phosphatase zeta 	2.18089E-24
Carbonic anhydrase 2	ENSP00000403028 <ul style="list-style-type: none"> known chromosome 	1.33133E-29
Carbonic anhydrase 2	ENSP00000143088 <ul style="list-style-type: none"> known chromosome CA-XII Carbonate dehydratase XIII Carbonic anhydrase 12 Carbonic anhydrase XIII Tumor antigen HGM-RCC-3.1.3 	3.70609E-38
Carbonic anhydrase 2	ENSP00000178638 <ul style="list-style-type: none"> known chromosome CA-XII Carbonate dehydratase XIII Carbonic anhydrase 12 Carbonic anhydrase XIII Tumor antigen HGM-RCC-3.1.3 	3.32903E-38
Carbonic anhydrase 2	ENSP00000209200 <ul style="list-style-type: none"> known chromosome CA-IV Carbonate dehydratase IV Carbonic anhydrase 4 Carbonic anhydrase IV 	6.28277E-30
Carbonic anhydrase 2	ENSP00000285379 <ul style="list-style-type: none"> known chromosome CA-II CA-C Carbonate dehydratase II 	1.44025E-153

Figure 3.7: Chemical Search results page

3.2.2 Protein-Ligand interactions Display

This page is found after performing a protein search in Discovery 2.0 and the tab will only appear if there is a BLAST or Domain match with at least one ChEMBL protein. The view is broken down into two displays (tabs), the first is the “By target” tab and the second is the “By ligand” tab. One function that can be carried out on either of the tabs is the “Get SDF” function that allows a user to pick compounds that he is interested in and save a .sdf file containing those compounds locally to perform further analysis. The function works slightly differently depending on whether it is triggered on the “By target” or the “By ligand” tabs. On the “By target” tab the user selects the protein target and all its associated compounds are added to the .sdf file and the “By ligand” tab uses the “Ligand efficiency index” plot to select compound subsets.

3.2.2.1 By Target

This display shows the results of performing a protein BLAST against the ChEMBL targets database, which is described in detail in the methods section. The user can at any point download an sdf file containing all or a subset of the molecules found, a subset can be selected by protein or by ligand and even using the ligand efficiency index to filter compounds. The results are ordered top down using the smallest E-value as an indicator of the best hit. There

The screenshot shows the DISCOVERY Malaria Drug Target Discovery System interface. The 'Protein-ligand interactions' tab is selected, and the 'By target' view is active. A table displays the following data:

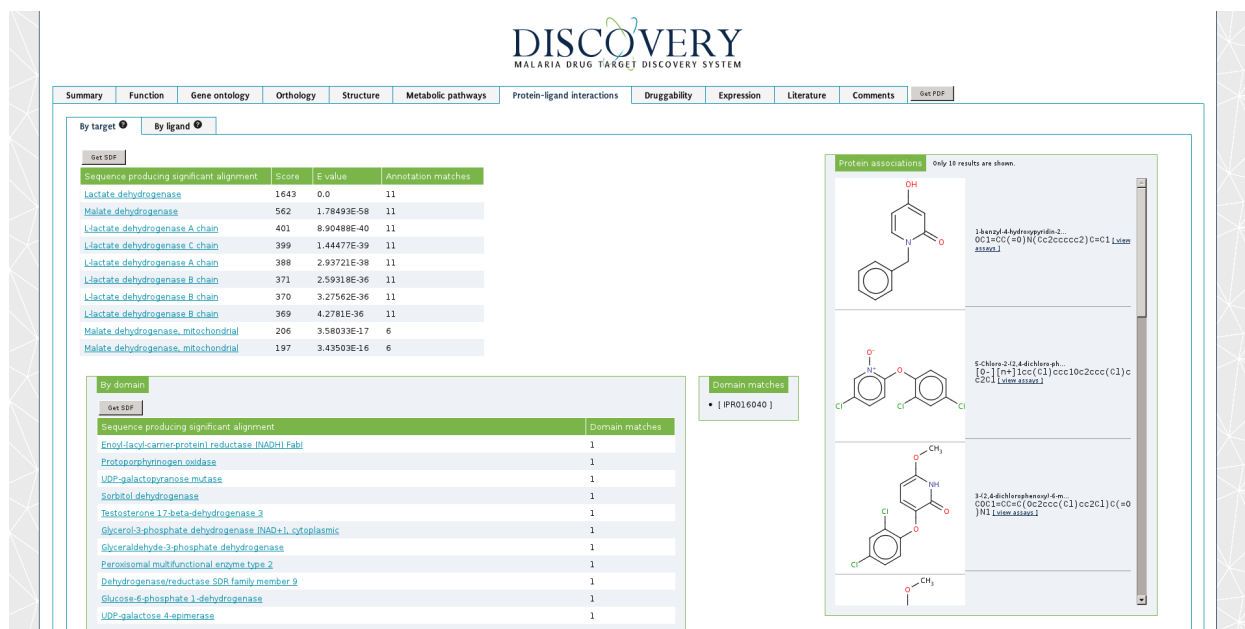
Sequence producing significant alignment	Score	E value	Annotation matches
Lactate dehydrogenase	1643	0.0	11

Below this table, a Smith-Waterman sequence alignment is shown between the query protein (Lactate dehydrogenase) and a ChEMBL target protein (Lactate dehydrogenase A chain). The matching regions are highlighted in green. To the right, the 'Protein associations' panel displays three entries, each with a 2D chemical structure, the ligand name, and its canonical SMILES string. The first entry is 2-amino-2-oxoacetic acid with SMILES NC(=O)C(=O)O. The second and third entries are 2-methoxy-4-(2-ethyl-...)-... and 2-cyano-4-(2-ethyl-...)-... with SMILES COc1ccc(CNC(=O)C)cc1 and C#Nc1ccc(CNC(=O)C)cc1.

Figure 3.8: The Protein-Ligand interactions tab (“By target”) after selecting BLAST hit.

are 4 columns of data displayed, they are the ChEMBL protein name, the alignment score, the E-value and the number of annotation matches the hit shares with the query protein. Under this list is a list of all ChEMBL target proteins that have matching domains, this list shows just two columns being the protein name as well as the number of shared domains. The domains were identified by using InterProScan as described in the methods section.

Figure 3.8 shows the result if the user selects a protein from the BLAST results section, a sub-window will appear directly below the entry in the list and a Smith-Waterman sequence alignment between the malaria and ChEMBL target protein is calculated and displayed with the matching area highlighted in green, this functionality is to allow for the user to see the actual alignment which will help determine the actual legitimacy of the hit, generally a researcher will look for alignments that have longer sections of conserved areas. All compounds associated with that ChEMBL target are then also displayed in a small display box that appears on the left. This Display box contains the 2D chemical structure, the ligand name and its canonical smiles string. Each entry also contains a link to a list of the assays that the compound is found in, which in turn also links back to the assay entry in the ChEMBL database. Any compounds that occur in the St Judes, Novartis Malaria Box and TCAMS



The screenshot shows the DISCOVERY Malaria Drug Target Discovery System interface. The 'By target' tab is selected, and the 'By domain' sub-tab is active. The main content area is divided into three sections:

- Sequence producing significant alignment:** A table listing protein sequences and their alignment statistics.

Sequence producing significant alignment	Score	E value	Annotation matches
Lactate dehydrogenase	1643	0.0	11
Malate dehydrogenase	562	1.78493E-58	11
Llactate dehydrogenase A chain	401	8.90489E-40	11
Llactate dehydrogenase C chain	399	1.44477E-39	11
Llactate dehydrogenase A chain	388	2.93721E-38	11
Llactate dehydrogenase B chain	371	2.59318E-36	11
Llactate dehydrogenase B chain	370	3.27562E-36	11
Llactate dehydrogenase B chain	369	4.2781E-36	11
Malate dehydrogenase_mitochondrial	206	3.58033E-17	6
Malate dehydrogenase_mitochondrial	197	3.43503E-16	6
- By domain:** A table listing domain matches.

Sequence producing significant alignment	Domain matches
Enoyl-acyl carrier protein reductase (NADH) FabI	1
Protoporphyrinogen oxidase	1
UDP-galactose 4-epimerase	1
Sorbitol dehydrogenase	1
Testosterone 17-beta-dehydrogenase_3	1
Glycerol-3-phosphate dehydrogenase (NAD+), cytoplasmic	1
Glyceralddehyde-3-phosphate dehydrogenase	1
Peroxisomal multifunctional enzyme type 2	1
Dehydrogenase/reductase SDR family member 9	1
Glucose-6-phosphate 1-dehydrogenase	1
UDP-galactose 4-epimerase	1
- Protein associations:** A section showing chemical structures and associated SMILES strings for three different compounds.
 - Structure 1: O=C1C=CC(=O)N(C1)Cc2ccccc2
 - Structure 2: Clc1ccc(Oc2ccc(Cl)cc2)cc1
 - Structure 3: COc1ccc(Oc2ccc(Cl)cc2)cc1

Figure 3.9: The Protein-Ligand interactions tab (“By target”) after selecting Domain hit.

data set will have the protein name highlighted in red.

Figure 3.9 shows the results if the user selects a protein from the “Domain matching” section, two data boxes will appear, the first is a box containing all the compounds associated with the matching protein identical to box that is displayed when a BLAST result is selected, the second is a list of all the shared domains.

3.2.2.2 By Ligand

This tab (Figure 3.10) displays all the ligands associated with the ChEMBL targets proteins that were identified via the protein blast or matching domains, all these compounds are accessible through the “By target” view, but if a user is more familiar with chemical compounds than proteins then this view is of significantly more value. The view is once again broken up into two lists, the first being all the compounds associated to proteins identified through BLAST and the second all the compounds associated to proteins that have matching domains. Both lists show the chemical name and have a link to view the chemical structure and chemical characteristics and a link to the ChEMBL database entry. What is added to this view is a scatter plot of the logP *vs.* Molecular weight of all the chemicals in the

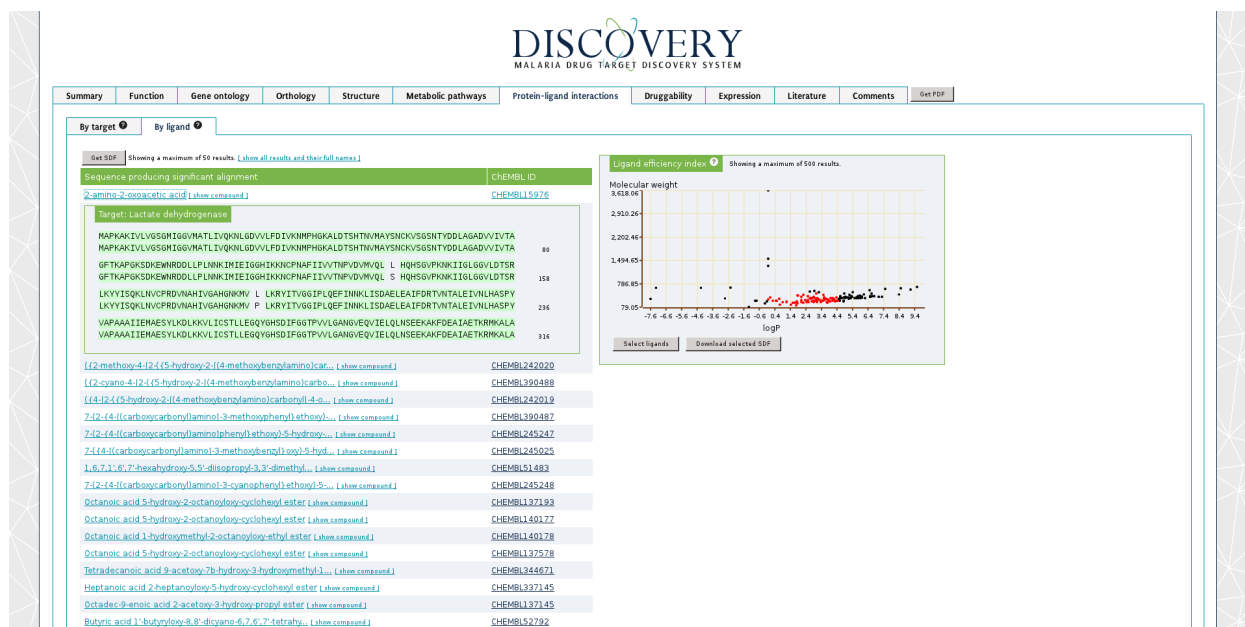


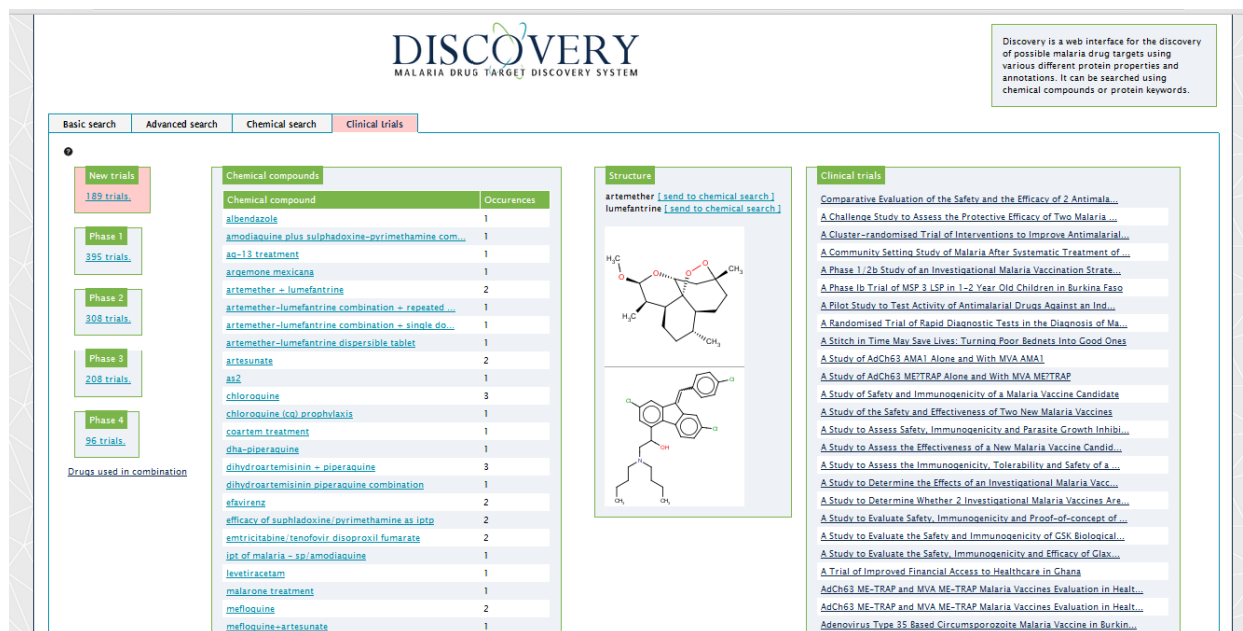
Figure 3.10: The Protein-Ligand interactions tab (“By ligand”) after selecting a compound.

lists, these are good indicators of ligand efficiency as hydrophobicity and molecular weight are large contributing factors for ligand binding. This has been termed the ligand efficiency index. The user is able to highlight a subset of molecules by selecting ranges that he finds acceptable. This was added to allow a user to filter the results to some degree, this filtering does not influence the compounds displayed but rather allows export of a filtered sdf file containing all the selected compounds.

Similar to the protein view if the user clicks on a compound from the BLAST results list the Smith-Waterman sequence alignment is displayed for the user to determine if the match is valid or not. If a user selects a compound from the matching domains list, the list of shared domains is displayed. The user can at any point download an sdf file containing all or a subset of the molecules found, a subset can be selected by protein or by ligand and even using the ligand efficiency index to filter compounds.

3.2.3 Clinical Trials

This page is a conglomeration of all trials registered at clinicaltrials.gov (“http://clinicaltrials.gov/”) involving malaria drugs being tested globally at various stages of development. The user can



DISCOVERY
MALARIA DRUG TARGET DISCOVERY SYSTEM

Discovery is a web interface for the discovery of possible malaria drug targets using various different protein properties and annotations. It can be searched using chemical compounds or protein keywords.

Basic search Advanced search Chemical search **Clinical trials**

New trials
189 trials

Phase 1
395 trials

Phase 2
308 trials

Phase 3
208 trials

Phase 4
96 trials

Drugs used in combination

Chemical compound	Occurrences
albendazole	1
amodiaquine plus sulphadoxine-pyrimethamine com...	1
aq-13 treatment	1
arcomone mexicana	1
artemether + lumefantrine	2
artemether-lumefantrine combination - repeated ...	1
artemether-lumefantrine combination - single do...	1
artemether-lumefantrine dispersible tablet	1
artesunate	2
as2	1
chloroquine	3
chloroquine (cq) prophylaxis	1
coartem treatment	1
dha-piperazine	1
dihydroartemisinin + piperazine	3
dihydroartemisinin piperazine combination	1
efaviranz	2
efficacy of sulphadoxine-pyrimethamine as iptp	2
emtricitabine/tenofovir disoproxil fumarate	2
ipt of malaria - sp/amodiaquine	1
levetiracetam	1
malarone treatment	1
mefloquine	2
mefloquine-artesunate	1

Structure
artemether [send to chemical search]
lumefantrine [send to chemical search]

Clinical trials
Comparative Evaluation of the Safety and the Efficacy of 2 Antimala...
A Challenge Study to Assess the Protective Efficacy of Two Malaria...
A Cluster-randomised Trial of Interventions to Improve Antimalarial...
A Community Setting Study of Malaria After Systematic Treatment of...
A Phase 1/2b Study of an Investigational Malaria Vaccination Strate...
A Phase Ib Trial of MSP 3 LSP in 1-2 Year Old Children in Burkina Faso
A Pilot Study to Test Activity of Antimalarial Drugs Against an Ind...
A Randomised Trial of Rapid Diagnostic Tests in the Diagnosis of Ma...
A Stitch in Time May Save Lives: Turning Poor Bednets Into Good Ones
A Study of AdCh63 AMA1 Alone and With MVA AMA1
A Study of AdCh63 METRAP Alone and With MVA METRAP
A Study of Safety and Immunogenicity of a Malaria Vaccine Candidate
A Study of the Safety and Effectiveness of Two New Malaria Vaccines
A Study to Assess Safety, Immunogenicity and Parasite Growth Inhibi...
A Study to Assess the Effectiveness of a New Malaria Vaccine Candid...
A Study to Assess the Immunogenicity, Tolerability and Safety of a ...
A Study to Determine the Effects of an Investigational Malaria Vac...
A Study to Determine Whether 2 Investigational Malaria Vaccines Are...
A Study to Evaluate Safety, Immunogenicity and Proof-of-concept of ...
A Study to Evaluate the Safety and Immunogenicity of GSK Biological...
A Study to Evaluate the Safety, Immunogenicity and Efficacy of Glax...
A Trial of Improved Financial Access to Healthcare in Ghana
AdCh63 ME-TRAP and MVA ME-TRAP Malaria Vaccines Evaluation in Healt...
AdCh63 ME-TRAP and MVA ME-TRAP Malaria Vaccines Evaluation in Healt...
Adenovirus Type 35 Based Circumsporozoite Malaria Vaccine in Burkin...

Figure 3.11: The Clinical trials tab.

at any point download an sdf file containing all or a subset of the molecules found, a subset can be selected by protein or by ligand and even using the ligand efficiency index to filter compound stages of development. Discovery 2.0 runs weekly queries against clinicaltrials.gov in order to update with recently registered trials, when a trial has been registered less than one month prior the Clinical trials tab on the start page is highlighted in red. On the clinical trials page the new trials are also accessible through a link that will also be highlighted red when it contains an entry. The page is primarily broken down into the 4 developmental phases, a link exists for each phase and when clicked it opens up a data box that lists all the compounds that are present in the clinicaltrials.gov data, along with the number of entries that drug occurs in the specific phase. The drugs are also categorized into trials where they occur in combination therapies, and each entry when clicked will display the chemical structure and a list of all the clinical trials linked back to the clinicaltrials.gov entries. At this point it is also possible to send the chemical structures to the sketch tool on the chemical search page and explore the possible protein targets.

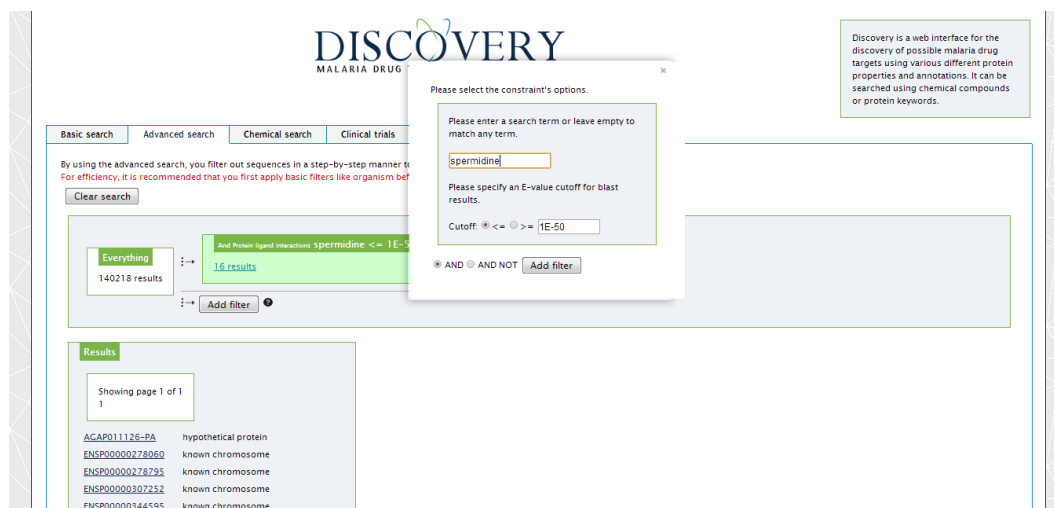


Figure 3.12: The advanced search feature showing a protein-ligand interactions filter.

3.2.4 Advanced Search

The advanced search is a powerful new feature released with the new Discovery system. This provides the user a platform through which he may provide logical filters to the main dataset of proteins in order to generate a subset that have the characteristics he may be interested in as illustrated in Figure 3.12. The example shows the results of providing a protein-ligand interaction filter using the keyword “spermidine”, this actually filters on the keyword in the protein name first then adds entries that have blast scores higher than the one specified in the filter to the list. The objective was to add more flexibility to the search to allow for additional proteins that may also interact with the same compounds. This is useful when one does not have an exact protein name or compound from which to begin a search with or is interested only in a particular property of a protein.

3.3 Data Statistics

The following are the basic statistics of the data of protein-ligand interactions presented in the Discovery 2.0 system.

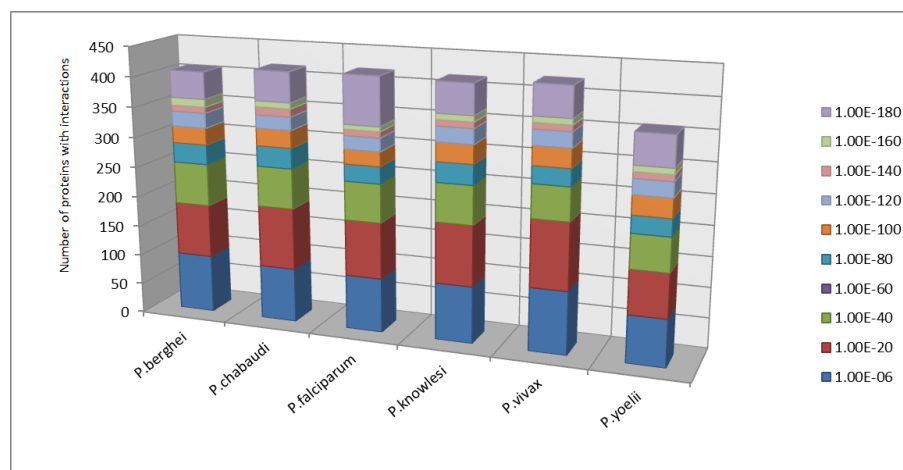


Figure 3.13: A stacked bar chart indicating the E-value of the best hit per *Plasmodium* protein that has at least one BLAST hit.

3.3.1 BLAST results

The number of *Plasmodium* proteins that have at least one blast hit against the ChEMBL targets database is 2,400. This can be broken down into the different species: *P.falciparum*: 422; *P.berghei*: 387; *P.chabaudi*: 391; *P.knowlesi*: 401; *P.vivax*: 411; *P.yoelii*: 388. Figure 3.13 shows a stacked bar indicating the number of proteins against the E-value of the best hit in the result set. The general trend with each species is that more than 50% of the matches have an E-value greater than $1e^{-80}$, and with a slightly inflated number of proteins with an E-value smaller than $1e^{-180}$ due to the exact matches that exist in the data, this being even more emphasized in *P. falciparum* due to it being the most studied species of *Plasmodium*. This chart indicates that there is a number of proteins with significantly strong BLAST hits to be considered good matches to associate ligand interactions. This number is for each of the species is still under 10% of the organisms proteins.

Figure 3.14 shows a stacked bar representing how many blast hits were found per protein, and clustering them in subgroups of 10. The trend shows what is to be expected, that majority of proteins have less than ten BLAST hit matches and represent proteins that are possibly less studied are more divergent than those that have many more hits. Those proteins with more than 10 BLAST hits likely represent proteins that are already well characterized and show stronger conservation between organisms.

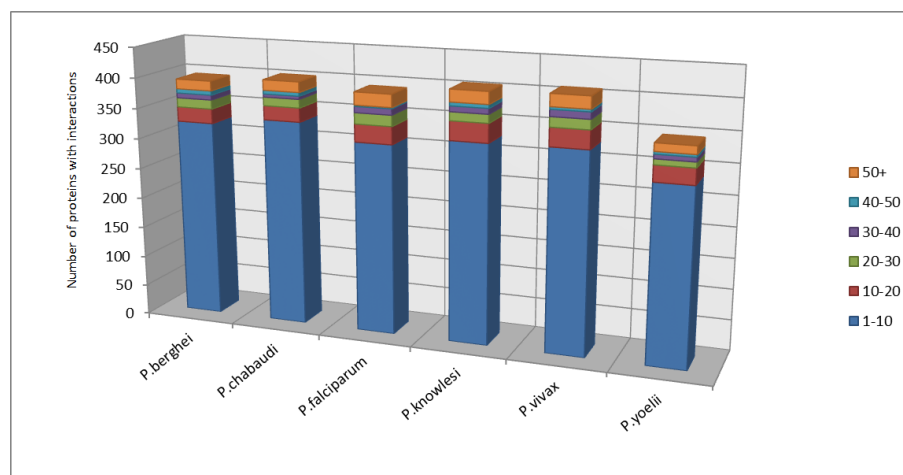


Figure 3.14: A stacked bar chart indicating the number of BLAST hits found per *Plasmodium* protein that has at least one domain hit

3.3.2 Domain results

The total number of proteins with domain matches from all species of malaria that are in Discovery 2.0 are 11,326. That can be broken down into the different species: *P. falciparum*: 1,996; *P. berghei*: 1,829; *P. chabaudi*: 1,847; *P. knowlesi*: 1,895; *P. vivax*: 1,940; *P. yoelii*: 1,819. Figure 3.15 shows the number of shared domains between the Plasmodium protein and the best ChEMBL targets match (ie the match containing the most shared domains). The results validate the reasonable assumption that there will be a significantly larger number of proteins that have matches below 5 domains. A protein match containing only one domain still carries the potential of having shared ligand activity, as functional domains play a strong role in protein activity. The goal of Discovery 2.0 in the instance of a protein that has only single domain match was to allow the possibility for the researcher to analyze that domain and interpret for himself if that domain will play a significant role in ligand activity which he can then deduce if the chemical association is valid or not. Further filtering of domains was deliberately not pursued to maximize the search potential of this approach.

Figure 3.16 shows the number of proteins with shared domains for each *Plasmodium* protein that has at least one match. The trend is once again as can be expected with many proteins having fewer than 30 matches. This is a good indication of how well documented a particular domain is which will be a good indicator of the likelihood of it being an already

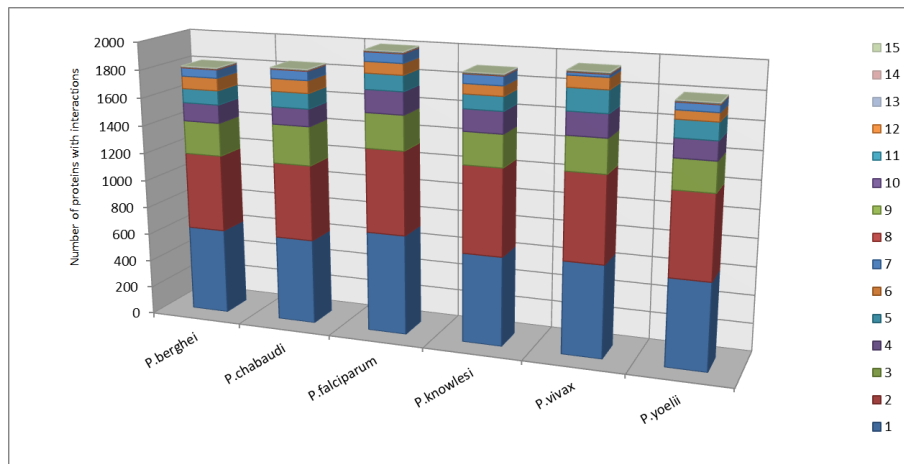


Figure 3.15: A stacked bar chart indicating the number of shared domains of the best hit

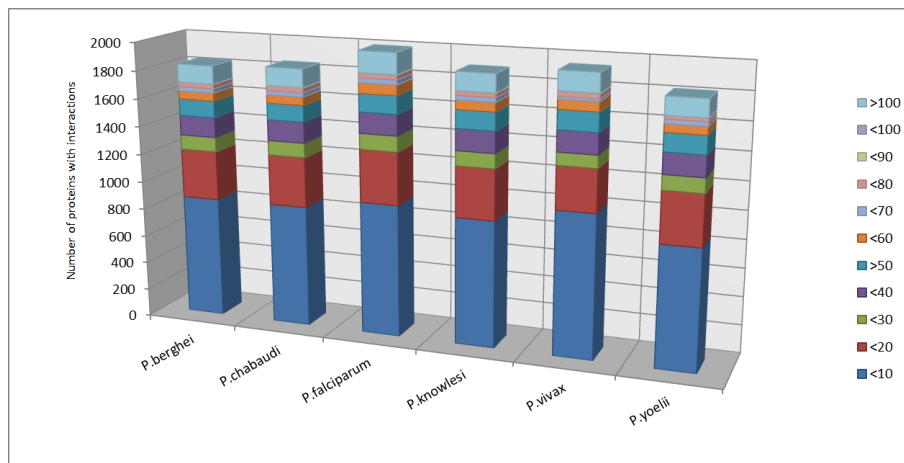


Figure 3.16: A stacked bar chart indicating the number of hits that have a shared domain for each *Plasmodium* protein

explored drug target. This deduction is due to the fact that the more matches a particular protein has the more times that particular set of domains has been through chemical assays which implies that it had been considered as a viable drug target candidate at some point. This does not indicate an increase in the probability of a valid protein-ligand interaction, but rather if the protein-ligand interaction is valid, it indicates the depth of information that is potentially available for further study.

3.4 Discussion

The above results illustrate the current scale of protein-ligand interaction prediction in Discovery 2.0 as well as provide a comparison between the old Discovery system and the new. The final result was a more efficient system that utilizes a significantly larger data source to make predictions against, this in itself led to more malaria proteins having associated ligands. The improvements to the interface were essential to being able to browse through this data in a logical manner due to the addition of domain based matches as well as the potential for some proteins to have significantly more hits. These improvements also extended the users ability to perform ligand based searching, for instance if a researcher has a series of molecules that he knows inhibit malaria but is not sure as to it's activity, he now has a decent chance of identifying a possible protein inter-actor for that protein. This was much less likely in the previous version of Discovery that only used the Drugbank database as a source.

Chapter 4

Validation

To measure the strength of Discovery 2.0 several case studies were performed on various proteins, and a comparison was made between the chemical inter-actors that were found in literature to interact with the proteins and the predictions made by the Discovery 2.0 system. It is important to note again that the ChEMBL database is built directly from literature. Due to the extremely large number of scientific publications it does not necessarily imply that the articles selected are as of yet contained within ChEMBL, however should the ChEMBL project continue the results will eventually be incorporated once they move onto various different journals. One goal in this section is to confirm whether Discovery 2.0 can correctly identify compounds already proven to inhibit *Plasmodium* proliferation, firstly using proteins that should have direct links to identical protein studies and secondly proteins that have high scoring matches and lastly proteins that have low scoring matches.

4.1 Case Studies

A series of proteins were chosen for case studies. The objective was to firstly measure the effectiveness of the system in identifying already known interactions of malaria proteins, then secondly to view the potential to expand on the already known set. Thirdly, to investigate the systems ability to find interactions where limited interaction data is available for the compound. Subsequently 2 compounds currently used as therapeutics are run through the

system in an attempt to see if the protein target can be identified.

4.1.1 Tanimoto Distance Significance

The following case studies are conducted using Tanimoto distances between molecules. Similarity scores of this kind can, however, be interpreted subjectively and deciding on what score is considered to be significant needs to be determined beforehand. This follows a similar procedure to that followed by Baldi *et al.* (2010) to try to determine what similarity score should be considered significant. To do this 40 compounds were randomly selected from the literature sets in the coming sections and their SMILES strings obtained. The tanimoto distance was then calculated between each compound and each entry in the ChEMBL database to determine the mean score and its distribution. This was performed on the UP bioinformatics server, the results collated and presented in Figure 4.1. Baldi *et al.* (2010) state that a similarity score needs to be taken in context of what type of data you are comparing and how it is being compared. In the case of Discovery where we are actually selecting compounds based on their associated protein distances and not directly based on the molecular distances of the compounds less stringency is required when evaluating the similarity scores. The mean value of these distributions was found to be 0.2202 with a standard deviation of 0.0785. Calculating the Z-score for a tanimoto distance of 0.4 one gets 2.290445, which can be computed into a p-value of 0.022 and can be considered statistically significant. This taken into account along with the displayed distributions a value of 0.4 will be stated as significant throughout the case studies. This value does not play a significant role in the analysis itself and merely helps with the interpretation of the results.

4.1.2 Lactate Dehydrogenase

Lactate dehydrogenase (LDH) was the first protein chosen for case study because of the wealth of knowledge already available on it's function, structure and it's significance in *Plasmodium falciparum* specifically. LDH is found in almost all animal tissues. The primary function of LDH is the catalysis of the interconversion of pyruvate and lactate combined

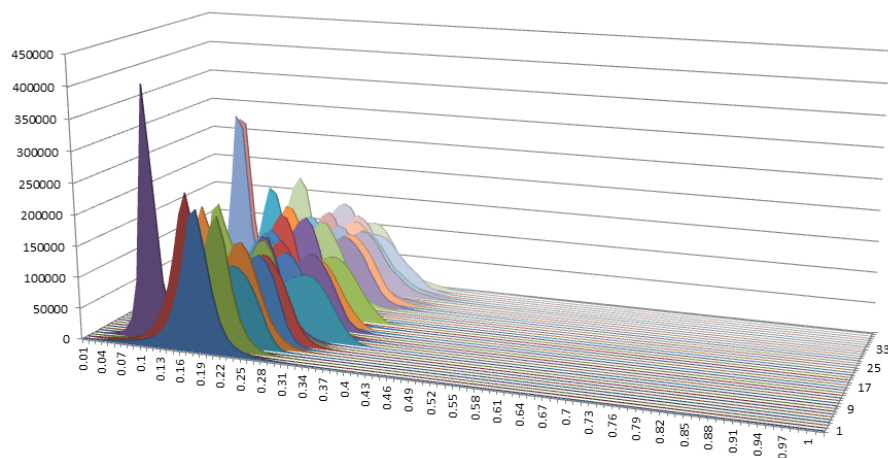


Figure 4.1: Tanimoto distance distributions of 40 randomly selected compounds, the x-axis shows the Tanimoto score, the y-axis the number of hits with that score and the z-axis is the arbitrary number representing the compound.

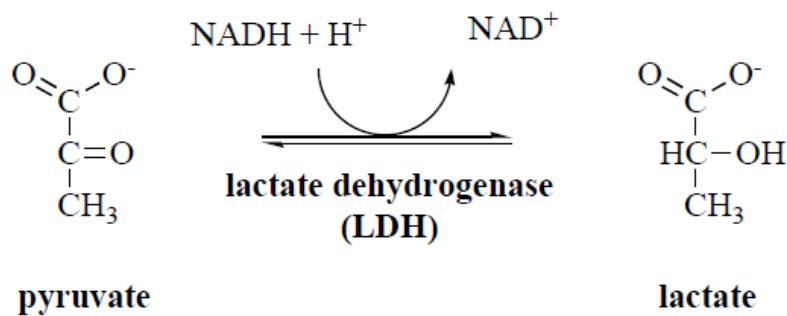


Figure 4.2: The reaction catalyzed by LDH

with the interconversion of NADH and NAD^+ . Under anaerobic conditions, such as events when there is a sudden demand for energy and low availability of oxygen, LDH converts pyruvate (the final step of glycolysis) to lactate the conversion is known as anaerobic homolactic fermentation. In doing so LDH allows the organism to overcome temporary anaerobic conditions by regenerating NAD^+ which is the electron acceptor during glycolysis and storing up lactate which is reconverted back to pyruvate when oxygen becomes available again (Everse and Kaplan, 1973).

LDH also performs the reverse reaction in gluconeogenesis during the Cori cycle, a metabolic pathway to produce glucose and thus ATP taking place primarily in the liver. This takes place during strenuous muscle activity when the blood glucose level decreases. In this situation the lactate that is produced in anaerobic conditions is converted back to pyruvate in the cytosol and internalized by the mitochondria. In this way gluconeogenesis also prevents lactic acidosis (Markert, 1984).

Plasmodium LDH (pLDH) is expressed at high levels in asexual stages of malaria parasites. pLDH activity is correlated with the level of parasitemia found in *in vitro* cultures of malaria and in the plasma of infected patients as determined by microscopy (Makler et al., 1993). pLDH isoforms are distinguishable from human isoforms on the basis of unique epitopes in pLDH and on enzymatic characteristics. Specifically pLDH has the ability to use the NAD analog 3 acetyl pyridine adenine dinucleotide (APAD) in the conversion of lactate to pyruvate (Jagt et al., 1981). Because of this feature pLDH can be easily distinguished in blood samples by measuring the conversion of APAD to APADH. This characteristic has also been used in the diagnosis of malaria because the turnover number of the pLDH in the presence of APAD is much greater than that of the human enzyme allowing samples to be easily distinguishable by measuring LDH activity in an assay (Hänscheid, 1999).

The protein information of plasmodium lactate dehydrogenase PF13_0141 has been summarized in 4.1.

Category	Type of Annotation	Annotation
Summary	aliases	L-lactate dehydrogenase
	sequence length	316
	identifiers	-PF13_0141 (plasmodb)
		-Q76NM3(Uniprot)
		-pfa:PF13_0141(KEGG gene)
Pubmed articles	-216	
Function	Families	-none
	Domains	-L-lactate/malate dehydrogenase (IPR001557)
		-Lactate dehydrogenase/glycoside hydrolase, family 4, C-terminal (IPR015955)
		-Lactate/malate dehydrogenase, C-terminal (IPR022383)
		-Lactate/malate dehydrogenase, N-terminal (IPR001236)
Gene Ontology	Molecular functions	GO:0000166 - nucleotide binding GO:0003824 - catalytic activity GO:0004459 - L-lactate dehydrogenase activity GO:0016491 - oxidoreductase activity GO:0016616 - oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
	Biological Processes	GO:0005975 - carbohydrate metabolic process GO:0044262 - cellular carbohydrate metabolic process GO:0055114 - oxidation-reduction process
Orthology	- <i>H.Sapiens</i> orthologs	-ENSP00000229319 -ENSP00000280704 -ENSP00000280706 -ENSP00000302393 -ENSP00000368722 -ENSP00000379385 -ENSP00000379386 -ENSP00000379516 -ENSP00000379518 -ENSP00000379524 -ENSP00000395337 -ENSP00000404535 -ENSP00000406172
	- <i>A.Gambiae</i> orthologs	-AGAP004880-PA -AGAP004880-PB -AGAP004880-PC
Metabolic pathways	KEGG	-Glycolysis / Gluconeogenesis -Cysteine and methionine metabolism -Pyruvate metabolism -Propanoate metabolism -Metabolic pathways -Biosynthesis of secondary metabolites
	MPMP	Established and putative Maurers clefts proteins S-Glutathionylated proteins Glycolysis Lactate dehydrogenase Total palmitome of <i>Plasmodium falciparum</i> Proteins targeted by the thioredoxin superfamily enzymes
	EC number	1.1.1.27 (L-lactate dehydrogenase)

Table 4.1: Summary of annotation data of L-lactate dehydrogenase (PF13_0141), excluding protein-ligand information.


Category	Result
Articles referenced for literature comparisons	Inhibitors of Lactate Dehydrogenase Isoforms and their Therapeutic Potentials. (Granchi et al., 2010) . Selective Inhibitors of Human Lactate Dehydrogenases and Lactate Dehydrogenase from the Malarial Parasite <i>Plasmodium falciparum</i> (Deck et al., 1998). Design, Synthesis, and Biological Evaluation of <i>Plasmodium falciparum</i> Lactate Dehydrogenase Inhibitors (Choi et al., 2007).
Number of ligands extracted from Literature	54
number of ligands found in Discovery	156; 144 from BLAST; 12 from 2 or more domain matches.
Venn Diagram	 <p>Blue: from BLAST matches. Green: from Domain matches with 2 or more domains. Red: Molecules extracted from literature.</p>

Table 4.2: A Summary of the protein-ligand interactions data of lactate dehydrogenase (PF13_0141).

4.1.2.1 Discovery 2.0 Compounds

A search was performed on the Discovery 2.0 start page using “lactate dehydrogenase” as a keyword search and the *Plasmodium falciparum* l-lactate-dehydrogenate (PF13_0141) was selected from the result set. After navigating to the protein-ligand interactions page, two separate sets of compounds were gathered. First, all compounds identified through BLAST matches were downloaded using the built-in “get sdf file” function. There were 10 matched sequences with the lowest E-value being $3.4 * 10^{-16}$, which resulted in an sdf containing 296 compounds. After all duplicates were removed we were left with 145 compounds. Second, a similar process was followed for the matched domains section. There were 8 unique entries that shared 5 domains, 1 that shared 2, and 85 unique entries that shared just a single domain, this gave us an sdf file containing 133 compounds, after duplicates were removed there were 105 compounds remaining.

These two sets were then screened individually against the compounds found in literature and each other for matching or similar hits, a match is categorized as something with a Tanimoto similarity score higher than 0.9 and a similar hit is a compound with a score higher than 0.4. The relevance of this is to illustrate that despite the matches not necessarily being exact the system is still able to find compounds with some degree of similarity. The diagram in Figure 4.2 shows a venn diagram explaining the different sets of compounds and the matches between them with the number of similar hits included in brackets. The intersection between all three areas was inconsequential thus not calculated. The screening was performed using the RDKit library which is an open-source cheminformatics and machine learning package. A series of python scripts were written to extract the SMILES strings from each of the sets, remove all duplicate records and convert the remainder into rdkit molecule objects that could then be used to calculate the tanimoto distance.

4.1.2.2 Literature Compounds

The compounds from the literature sources were selected by performing a search on Google Scholar using the keywords “Plasmodium Lactate dehydrogenase inhibitors” and 3 articles

that referred to ligand inhibition of pLDH were found and selected. All compound structures found in the articles were run through PubChem to verify their existence and gather additional information on the molecule, in particular the SMILES string which is used to elucidate the Tanimoto distance (Wang et al., 2009). All compounds were then summarized into a single smiles (.smi) file.

4.1.2.3 Remarks

In this particular example the Discovery 2.0 system successfully identifies a significant number of the compounds collected from literature, this is not too surprising due to the depth of information available for LDH in literature and thus the prominence of ligand information available in ChEMBL. The BLAST searching was more successful at finding exact matches than the compounds found through shared domains. 2 compounds were found matching in all three sets, which were carbamoylformic acid also called oxalamic acid and 2,3-dihydroxy-6,7-dimethyl-4-(propan-2-yl)naphthalene-1-carboxylic acid, when searching for associated targets both correctly detected L-lactose dehydrogenase as predicted targets. Fourteen compounds were matched between the compounds from literature and those through BLAST results. The remainder of the compounds found with proteins with two or more domains were also found in the BLAST results, this is to be expected as the likelihood is high that those domains are binding domains that matched.

4.1.3 Dihydroorotate dehydrogenase

During the erythrocytic stage of the *Plasmodium* parasite, it undergoes active division which results in an large increase in the need for nucleic acids which are required for DNA, RNA, glycoprotein and phospholipid biosynthesis. Primates have a pathway for the salvage of pyrimidines that is absent in plasmodium, making *Plasmodium* entirely reliant on the *de novo* production of pyrimidines pathway this pathway is thus seen as a good target for drug development. A series of enzymes have been identified to be involved in the pyrimidine biosynthesis pathway including carbamoyl phosphate synthase, aspartate carbamoyltrans-

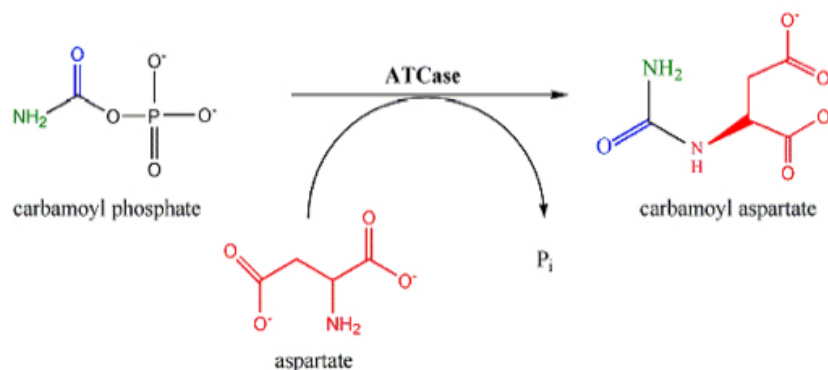


Figure 4.3: Reaction catalyzed by DHOD (Patel et al., 2008).

ferase, dihydroorotate, dihydroorotate dehydrogenase, orotate phosphoribosyl transferase, and orotidine 50-phosphate decarboxylase (Reyes et al., 1982, Jones, 1980). The focus in this case study is dihydroorotate dehydrogenase (DHOD) which is a mitochondrially localized flavozyme and functions in the pathway mentioned. DHOD catalyzes the oxidation of L-dihydroorotate (l-DHO) to orotate as part of the fourth and rate-limiting step of the *de novo* pyrimidine biosynthetic pathway as shown in Figure 4.3. Dihydroorotate dehydrogenase (DHODH) is a well-known protein target for *P.falciparum*. DHODH forms part of the pyrimidine biosynthesis pathway, by facilitating the conversion of L-dihydroorotate (DHO) to orotate (ORO) (Jones, 1980). Studies have shown that this enzyme possesses two distinct binding sites, respectively for DHO/ORO and ubiquinone (Davis et al., 1996). Oxidation of DHO to ORO is the rate-limiting step of the whole pyrimidine biosynthetic pathway.

DHOD was selected as a case study protein because it has various hits in the Discovery protein-ligand search results and is not currently an exploited anti-malarial.

4.1.3.1 Discovery 2.0 Compounds

A search was performed on the Discovery 2.0 start page using “Dihydroorotate dehydrogenase” as a keyword search and the *Plasmodium falciparum* dihydroorotate dehydrogenase, mitochondrial precursor (PFF0160c) was selected from the result set. Subsequently the download sdf file feature was used to gather a list of all compounds identified. These were then screened individually against the known compounds and the matching or significant

Category	Type of Annotation	Annotation
Summary	aliases	dihydroorotate dehydrogenase, mitochondrial precursor DHOdehase Dihydroorotate dehydrogenase (quinone), mitochondrial Dihydroorotate oxidase
	sequence length	54
	identifiers	-PFF0160c (plasmodb)
		-Q08210(Uniprot)
		-pfa:PFF0160c(KEGG gene)
Pubmed articles	-54	
Function	Domains	-DHODEHASE_2 (IPR001295)
		-DHO_dh (IPR012135)
		-Aldolase-type TIM barrel (IPR013785)
		-Dihydroorotate dehydrogenase, class 2 (IPR005719)
Gene Ontology	Molecular functions	GO:0003824 - catalytic activity GO:0004152 - dihydroorotate dehydrogenase activity GO:0016491 - oxidoreductase activity
	Cellular Components	GO:0005739 - mitochondrion GO:0005743 - mitochondrial inner membrane GO:0016020 - membrane GO:0016021 - integral to membrane
	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	-ENSP00000219240
	- <i>A. gambiae</i> orthologs	-AGAP002037-PA
	- <i>P. chambaui</i>	PCHAS_010280
	- <i>P. berghei</i>	-PBANKA_010210
	- <i>P. knowlesi</i>	-PKH_114660
	- <i>P. vivax</i>	-PVX_113330
	- <i>P. yoelii</i>	-PY02580
Metabolic pathways	KEGG	-Pyrimidine metabolism -Metabolic pathways
	EC number	1.3.98.1 (dihydroorotate dehydrogenase (fumarate))

 Table 4.3: Summary of annotation data of *Plasmodium falciparum* Dihydroorotate dehydrogenase (PFF0160c), excluding protein-ligand information.

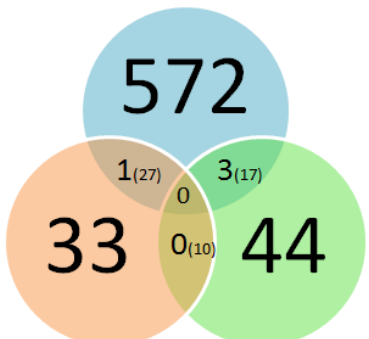
Category	Result
Articles referenced for literature comparisons	<p>New inhibitors of dihydroorotate dehydrogenase (DHODH) based on the 4-hydroxy-1,2,5-oxadiazol-3-yl (hydroxyfurazanyl) scaffold (Lolli et al., 2012).</p> <p>Identification and Characterization of Small Molecule Inhibitors of <i>Plasmodium falciparum</i> Dihydroorotate Dehydrogenase (Patel et al., 2008).</p> <p>Subset of 10 inhibitors described in patent Thunuguntla (2010).</p> <p>Small molecule inhibitors of <i>Plasmodium falciparum</i> dihydroorotate dehydrogenase (Bastos, 2011).</p>
Number of ligands extracted from Literature	34
number of ligands found in Discovery	623, 47 from domain matches, 576 from BLAST matches
Venn Diagram	 <p>Blue: from BLAST matches. Green: from Domain matches with 2 or more domains. Red: Molecules extracted from literature. The numbers in brackets represent matches with a 0.4 cut-off.</p>

Table 4.4: A Summary of the protein-ligand-interactions data of lactate dehydrogenase (PF13_0141).

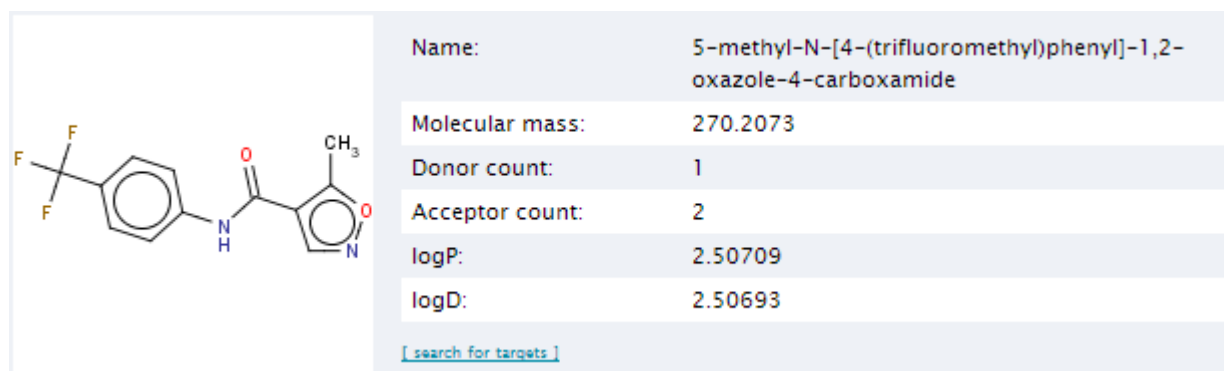


Figure 4.4: Leflunomide was a compound that was in the literature set and identified as a match in Discovery 2.0

hits are shown in a Venn diagram in Table 4.4. The diagram shows that the number of matches to the literature set were low with only one molecule found matching between the literature results and the blast results. However, when the Tanimoto score threshold for a hit is reduced to 0.4, the number of matches with the blast set goes up to 27 and the domains set goes up to 10. The point of making this comparison is to illustrate that some similarities are still present between the different subsets. The compound that was found to match was leflunomide represented in Figure 4.4 which is a well known inhibitor of DHODH, it is used as the base molecule by Lolli *et al.* (2012) which is one of the data sources used to compile the literature set.

4.1.3.2 Literature Sources

Patel *et al.* (2008) identified and characterized DHOD inhibitors using target-based HTS to identify chemical start points for drug development. They screened compounds from the Genzyme Corp small molecule library which comprised of 208,000 diverse, commercially available molecules. Of these 698 compounds were found that inhibited pDHOD at >70% at a concentration of 10 μ m. These compounds were then re-screened and 55 compounds were found that had inhibition of >50% at 1 μ m. Dose-effect curves identified 38 compounds with submicromolar IC₅₀'s. From the 38 pDHOD inhibitors that were evaluated for antimalarial efficacy using *P. falciparum* 3D7 as the test strain, five compounds were found to have submicromolar IC₅₀ values, they were subsequently tested for inhibitory activity against

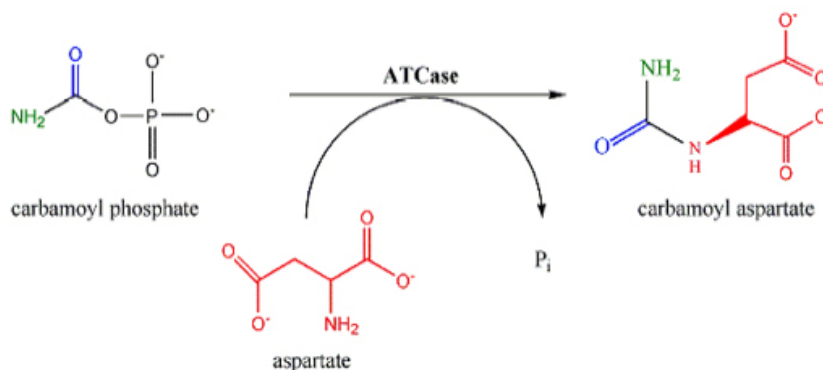


Figure 4.5: Reaction catalyzed by ATCase

drug resistant strains HB3 and Dd2. These five compounds were the first compounds taken for the literature set.

To expand the set of known inhibitors a subset of 10 compounds from patent data was taken from patent documents published in 2010 (Thunuguntla, 2010) which specify a list of characterized inhibitors of DHODH. And a further 5 were selected from another patent also describing DHODH inhibitors (Bastos, 2011).

4.1.4 Aspartate carbamoyltransferase

Aspartate carbamoyltransferase (ATCase) is another protein in the *de novo* biosynthesis pathway. It forms the first step and catalyzes the formation of phosphate and N-carbamoyl-L-aspartate from carbamoyl phosphate and L-aspartate in pyrimidine biosynthesis as illustrated in Figure 4.5.

4.1.4.1 Discovery 2.0 Compounds

A search was performed on the Discovery 2.0 start page using “Aspartate carbamoyltransferase” as a keyword search and the *Plasmodium falciparum* aspartate carbamoyltransferase (MAL13P1.221) was selected from the result set. Subsequently the download sdf file feature was used to gather a list of all compounds identified. These were then screened individually against the known compounds and the matching or significant hits are shown in the figure in Table 4.6. The screening was performed as described in the same way as described for the

Category	Type of Annotation	Annotation
Summary	aliases	aspartate carbamoyltransferase
	sequence length	375
	identifiers	-MAL13P1.221 (plasmodb)
		-Q8IDP8(Uniprot)
	-pfa:MAL13P1.221(KEGG gene)	
Pubmed articles		-2
Function	Families	-none
	Domains	-Aspartate/ornithine carbamoyltransferase (IPR006130)
		-Aspartate carbamoyltransferase (IPR002082)
		-Aspartate/ornithine carbamoyltransferase, carbamoyl-P binding (IPR006132)
	-Aspartate/ornithine carbamoyltransferase, Asp/Orn-binding domain (IPR006131)	
Gene Ontology	Molecular functions	GO:0004070 - aspartate carbamoyltransferase activity
		GO:0016597 - amino acid binding
		GO:0016740 - transferase activity
GO:0016743 - carboxyl- or carbamoyltransferase activity		
	Cellular Components	GO:0020011 - apicoplast
	Biological Processes	GO:0006207 - 'de novo' pyrimidine nucleobase biosynthetic process GO:0006520 - cellular amino acid metabolic process
Orthology	- <i>H. sapiens</i> orthologs	-ENSP00000405416
	- <i>P. chambaudi</i>	PCHAS_136230
	- <i>P. berghei</i>	-PBANKA_135770
	- <i>P. knowlesi</i>	-PKH_120960
	- <i>P. vivax</i>	-PVX_083135
	- <i>P. yoelii</i>	-PY06210
Metabolic pathways	KEGG	-Pyrimidine metabolism -Alanine, aspartate and glutamate metabolism -Metabolic pathways
	MPMP	-Asparagine and Aspartate metabolism -Nuclear genes with apicoplast signal sequences -Total palmitome of <i>Plasmodium falciparum</i> -Pyrimidine metabolism
	Reactome	carbamoyl phosphate + ornithine => citrulline + orthophosphate
	EC number	2.1.3.2


Category	Result
Articles referenced for literature comparisons	Aspartate carbamoyltransferase of <i>Plasmodium falciparum</i> as a potential drug target for designing anti-malarial chemotherapeutic agents (A.Banerjee et al., 2012).
Number of ligands extracted from Literature	10
number of ligands found in Discovery	11, 3 from domain matches, 8 from BLAST matches.
Venn Diagram	 <p>Blue: from BLAST matches. Green: from Domain matches with 2 or more domains. Red: Molecules extracted from literature. The numbers in brackets represent matches with a 0.4 cut-off.</p>

Table 4.6: A summary of the protein-ligand-interactions data of aspartate carbamoyltransferase (MAL13P1.221).

LDH case study.

4.1.4.2 Literature compounds

Banjeree *et al.* (2012) characterized aspartate carbamoyltransferase *in silico*, they derived the tertiary (3D) structure of the enzyme by using the structure of aspartate carbamoyltransferase of *Pyrococcus abyssi* (PDB ID: 1ML4) as template by comparative modeling and validated by various structural quality validation tools. Once the model was found to be stable in simulations a number of inhibitor molecules were docked to the models active site and the binding affinities recorded. Only 10 molecules were recorded in the publication

4.1.4.3 Remarks

The number of matches found in this case study was very small with only one match existing between the compound that was found in the literature and those identified through blast. The compound that was found to match is presented along with the BLAST hit result in Figure 4.6 the Tanimoto distance between the two molecules was 0.637. On visual comparison the difference is the positional change of the COOH group and a missing amine group. When cross referencing this molecule back through the Discovery 2.0 chemical search feature (Figure 4.7) it is found that the molecule is 2-Amino-5-(2-phosphono-acetylamino)-pentanoic acid. The molecule is described in 16 assays in ChEMBL (ChEMBL1160567), 3 Binding and 13 Functional, on the protein ornithine transcarbamoylase which is the blast hit protein. The results in the figure indicate that the blast e-value is 1.93E-15 which is not a particularly strong hit however the two proteins share 3 domains. This information adds confidence that this compound will bind to ATCase. The drug identified also has a link to Drugbank3.0 (DB02011 or EXPT02497) which has an entry marking this compound as experimental, and it does not contain the typical depth of information as a regular entry in Drugbank does.

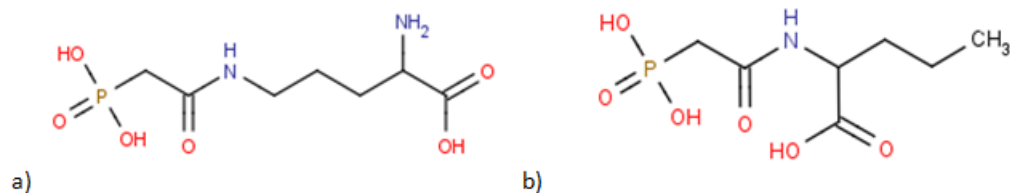

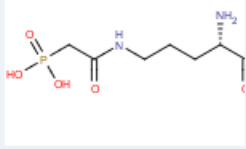


Figure 4.6: The 2D structure of the molecule found in Discovery(a) and the one predicted through docking methods and found in literature(b).





Name: (2S)-2-amino-5-(2-phosphonoacetamido)pentanoic acid

Molecular mass: 254.1776

Donor count: 5

Acceptor count: 7

logP: -4.25111

logD: -7.51105

[\[search for targets\]](#)

Only showing the results of a maximum of 5 chemical targets.

Chemical target	Associated protein	E value
Ornithine carbamoyltransferase, mitochondrial	ENSP00000405416 novel chromosome	2.26353E-13
Ornithine carbamoyltransferase, mitochondrial	ENSP00000384510 novel chromosome	1.17491E-19
Ornithine carbamoyltransferase, mitochondrial	ENSP00000264705 <ul style="list-style-type: none"> known chromosome Aspartate carbamoyltransferase CAD protein DE Includes DE Includes DE Includes Dihydroorotase Glutamine-dependent carbamoyl-phosphate synthase 	1.18787E-19
Ornithine carbamoyltransferase, mitochondrial	ENSP00000039007 <ul style="list-style-type: none"> known chromosome OTCase Ornithine carbamoyltransferase, mitochondrial Ornithine transcarbamylase 	0.0
Ornithine carbamoyltransferase, mitochondrial	AGAP000300-PA hypothetical protein	3.05003E-19
Ornithine carbamoyltransferase, mitochondrial	MAL13P1.221 aspartate carbamoyltransferase	1.93609E-15

Figure 4.7: Cross-Reference of matching ligand in Aspartate carbamoyltransferase

4.1.5 Sulfadoxine

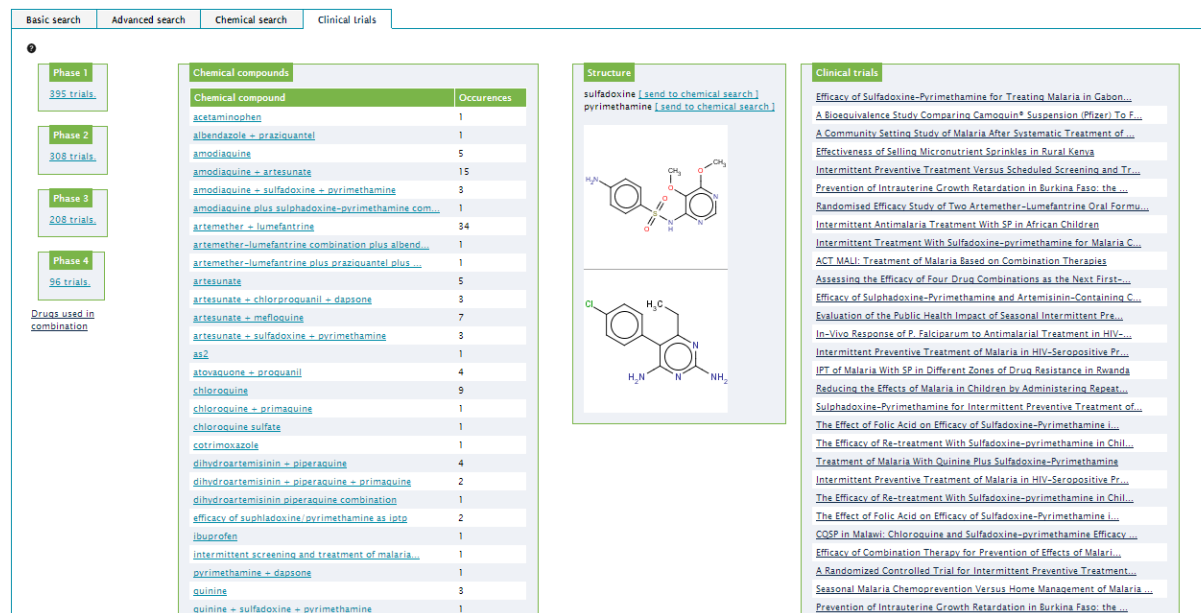
In this case study the intention is to test out the Discovery chemical search function effectiveness, the simplest way to do this is take one of the currently used anti-malarial drugs that have a known target and run it through the chemical search feature. In this case study we accessed the clinical trials page shown in Figure 4.8 to search for compounds that are currently being going through stage 4 trials because phase 4 trial drugs are already commercially available and being used to treat malaria infection. Sulfadoxine was chosen from the available list and the compound sent to the chemical search page and an exact match query performed, the result is shown in Figure 4.9. When selecting the “search for targets button” a long list of possible protein targets is listed, when specifically looking for Plasmodium proteins you find “PF08_0095” which is the PlasmoDB id for Dihydropteroate synthase which is the protein believed to be inhibited by sulfadoxine (Brooks et al., 1994, Triglia and Cowman, 1994). This was verified by running the query in reverse by selecting the protein search page and entering “PF08_0095” and searching through the protein-ligand interactions section and finding sulfadoxine as well as similar molecules in the results.

4.1.6 Pyrimethamine

In a similar fashion to Sulfadoxine, Pyrimethamine was also ported from the clinical trials page to the chemical search feature, and an exact search conducted. The list of interactors is significantly shorted to Sulphadoxine and the system correctly identifies PFD0830w bi-functional dihydrofolate reductase-thymidylate synthase(DHFR) as one of the possible interactors. This is a well studied interaction and Pyrimethamine is a proven inhibitor of DHFR (Bzik et al., 1987).

4.2 Discussion

The above case studies illustrate the accuracy that Discovery has in predicting real protein-ligand interactions. It also clearly illustrates how the predictive ability varies between proteins



Basic search | **Advanced search** | **Chemical search** | **Clinical trials**

Phase 1
395 trials

Phase 2
308 trials

Phase 3
208 trials

Phase 4
96 trials

Drugs used in combination

Chemical compounds

Chemical compound	Occurrences
acetaminophen	1
albendazole + praziquantel	1
amodiaquine	5
amodiaquine + artesunate	15
amodiaquine + sulfadoxine + pyrimethamine	3
amodiaquine plus sulphadoxine-pyrimethamine com...	1
artemether + lumefantrine	34
artemether-lumefantrine combination plus albed...	1
artemether-lumefantrine plus praziquantel plus ...	1
artesunate	5
artesunate + chloroquinil + dapsone	3
artesunate + mefloquine	7
artesunate + sulfadoxine + pyrimethamine	3
azs	1
atovaquone + proquanil	4
chloroquina	9
chloroquina + primaquina	1
chloroquina sulfate	1
cotrimoxazole	1
dihydroartemisinin + piperaquine	4
dihydroartemisinin + piperaquine + primaquine	2
dihydroartemisinin piperaquine combination	1
efficacy of sulphadoxine-pyrimethamine as iptp	2
ibuprofen	1
intermittent screening and treatment of malaria...	1
pyrimethamine + dapsone	1
quinine	3
quinine + sulfadoxine + pyrimethamine	1

Structure

[sulfadoxine \[send to chemical search\]](#)
[pyrimethamine \[send to chemical search\]](#)

Clinical trials

[Efficacy of Sulfadoxine-Pyrimethamine for Treating Malaria in Gabon...](#)
[A Bioequivalence Study Comparing Camoquin* Suspension \(Pfizer\) To F...](#)
[A Community Setting Study of Malaria After Systematic Treatment of ...](#)
[Effectiveness of Selling Micronutrient Sprinkles in Rural Kenya](#)
[Prevention of Intrauterine Growth Retardation in Burkina Faso: the ...](#)
[Intermittent Preventive Treatment Versus Scheduled Screening and Tr...](#)
[Prevention of Intrauterine Growth Retardation in Burkina Faso: the ...](#)
[Randomised Efficacy Study of Two Artemether-Lumefantrine Oral Formu...](#)
[Intermittent Antimalarial Treatment With SP in African Children](#)
[Intermittent Treatment With Sulfadoxine-pyrimethamine for Malaria C...](#)
[ACT MALI: Treatment of Malaria Based on Combination Therapies](#)
[Assessing the Efficacy of Four Drug Combinations as the Next First...](#)
[Efficacy of Sulphadoxine-Pyrimethamine and Artemisinin-Containing C...](#)
[Evaluation of the Public Health Impact of Seasonal Intermittent Pra...](#)
[In-Vivo Response of P. Falciparum to Antimalarial Treatment in HIV...](#)
[Intermittent Preventive Treatment of Malaria in HIV-Seropositive Pr...](#)
[IPT of Malaria With SP in Different Zones of Drug Resistance in Rwanda](#)
[Reducing the Effects of Malaria in Children by Administering Repeat...](#)
[Sulphadoxine-Pyrimethamine for Intermittent Preventive Treatment of ...](#)
[The Effect of Folic Acid on Efficacy of Sulfadoxine-Pyrimethamine I...](#)
[The Efficacy of Re-treatment With Sulfadoxine-pyrimethamine in Chil...](#)
[Treatment of Malaria With Quinine Plus Sulfadoxine-Pyrimethamine](#)
[Intermittent Preventive Treatment of Malaria in HIV-Seropositive Pr...](#)
[The Efficacy of Re-treatment With Sulfadoxine-pyrimethamine in Chil...](#)
[The Effect of Folic Acid on Efficacy of Sulfadoxine-Pyrimethamine I...](#)
[COGP in Malawi: Chloroquine and Sulfadoxine-pyrimethamine Efficacy ...](#)
[Efficacy of Combination Therapy for Prevention of Effects of Malari...](#)
[A Randomized Controlled Trial for Intermittent Preventive Treatment...](#)
[Seasonal Malaria Chemoprevention Versus Home Management of Malaria ...](#)
[Prevention of Intrauterine Growth Retardation in Burkina Faso: the ...](#)

Figure 4.8: Clinical-trials showing sulfadoxine-pyrimethamine combination. This snapshot shows all clinical trial being conducted on this specific combination of drugs. It also allows the user the “send to chemical search” link which navigate the user to the chemical search page and populates the selected molecule into the Marvin sketch plug-in

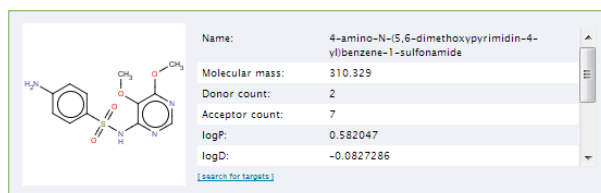


Figure 4.9: Sulfadoxine as viewable via the Discovery2.0 interface

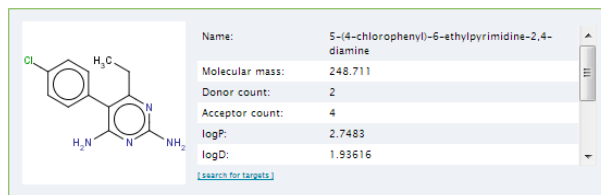


Figure 4.10: Pyrimethamine as viewable via the Discovery2.0 interface.

that have been well studied against those that have not, as it highlights the bias to gather more data on already identified targets as opposed to performing more broad studies. This discrepancy may be addressed in future as ChEMBL incorporates larger data sets and more large scale experiments are published and become publicly available.

Discovery does however still manage to successfully find matches even in such cases where little data is available, which was one of the main objectives. Discovery is also able to successfully identify a protein target of a chemical using the chemical search tool. This may potentially be one of the more interesting features as it may lead to the identification of novel biological processes or pathways in the *Plasmodium* parasite that were previously unknown.

Chapter 5

Concluding Discussion

The Discovery 2.0 system has successfully incorporated the ChEMBL database to be the primary source of bio-active molecule information. By incorporating the ChEMBL database into Discovery we have increased the available bio-active molecules from 6811 (Drugbank 3.0) to 1,324,941 distinct compounds and the number of associated protein targets from 4294 to 9,844 as of ChEMBL_16. With the continuation of the ChEMBL project and the Discovery system setup to automatically incorporate new versions of the ChEMBL database, we will have a continuously expanding data source that is externally maintained and funded.

This addition allows Discovery to leverage bio-active data of other organisms to increase our understanding of the malaria parasite. This in combination with the high quantity of data that is hosted within Discovery enables users to utilize a single source to carry out a their malaria data mining needs.

In the case studies we evaluate the accuracy by using literature studies and comparing known interactions with those predicted in Discovery. It does perform well in this regard, however because there is no way to evaluate putative interactions on undocumented proteins and performing the test screens was not in the scope of this study it was not possible to evaluate the error rate of the hits.

Discovery has been upgraded both in technical design and available features. With the conversion from the python platform to Java and the rework done to the database design, the web interface and data queries are notably quicker than Discovery 1.0 and the interface

allows for information to be presented to the user in a far more dynamic way. Discovery 2.0 has also been updated to automatically collect the most recent available data from the various sources to ensure that it is always up to date and populated with the most accurate information available. Discovery allows researchers to interact with malaria data from either chemical or protein data as start points, it also allows gathering of batch results based on logical filters via the advanced search feature. This provides a platform for researchers to use to quickly gather large quantities of relevant biological data about malaria.

Discovery 2.0 provides a logical interface that can be used to search for and gather relevant biological data for use of drug target and lead compound selection, Discovery is also able to predict possible interacting ligands and make that list available to the user for download. Such a list can be used for screening purposes or even pharmacophore identification.

Comparing Discovery to another similar malaria centered web resources, one being “PlasmoDB” (Aurrecoechea et al., 2009). Currently PlasmoDB does not have any linked protein-ligand information with it’s system, and despite that you are able to search for compounds there is no way to link them to specific proteins in their current system. PlasmoDB does however contain a much broader set of data types that include SNPs, Mass spec data, and even population data for malaria parasites, all of which are very useful and complimentary to Discovery.

Another such resource is TDRTargets (Magariños et al., 2012) which implemented a similar approach to gathering protein-ligand interactions, they also utilize the ChEMBL database, and use two methods of identifying interactions, firstly through orthology and secondly using BLAST. They have applied this to multiple species not just *Plasmodium* and have developed a weighting algorithm to act as a form of advanced searching that allows the customized ranking of results. The greatest difference between Discovery and TDRTargets currently is Discovery’s focus on the malaria parasite, and in context of protein-ligand interactions prediction using the domain matching approach Discovery has a greater number malaria proteins that now have associated chemical data. The quality of these hits may not always be that high due to the bias in research mentioned but it does however provide a potential

new insight into a proteins activity.

Discovery is used by searching through the biological data of the malaria proteins based on what the user perceives as the strongest describing factors of a good protein target, the factors can be: specific times of expression of a protein; the metabolic pathway in which it functions; its orthology with human proteins; or with the new functionality added whether or not protein-ligand interactions exist based on current knowledge. Using a broad range of characteristics it becomes easier to discern between good targets and bad and has the potential of saving precious time and resources by not pursuing a target that can be invalidated *in silico*.

In conclusion the Discovery resource is available at “<http://discovery.bi.up.ac.za/>” and is ready to be used in the effort to find alternative therapeutics to treat malaria infection. Its incorporation of various data types and external links to various external resources make it a particularly useful platform to search through the malaria related biological information.

Bibliography

A.Banerjee, N.Arora, and U.Murty. Aspartate carbamoyltransferase of *Plasmodium falciparum* as a potential drug target for designing anti-malarial chemotherapeutic agents. *Medicinal Chemistry Research*, 21:1–14, 2012. ISSN 1054-2523. URL <http://dx.doi.org/10.1007/s00044-011-9757-3>. 10.1007/s00044-011-9757-3.

A.Bender and R.C.Glen. Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem*, 2(22):3204–3218, Nov 2004. doi: 10.1039/B409813G. URL <http://dx.doi.org/10.1039/B409813G>.

A.Bender, D.W.AnsongYoung, J.L.Jenkins, M.Serrano, D.Mikhailov, P.A.Clemons, and J.W.Davies. Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb Chem High Throughput Screen*, 10(8):719–731, Sep 2007.

S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

T. K. Attwood, P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, and C. Zygouri. Prints and its automatic supplement, preprints. *Nucleic Acids Res*, 31(1):400–402, Jan 2003.

C. Aurrecochea, J.Brestelli, B.P. Brunk, J. Dommer, S. Fischer, B Gajria, X Gao, A Gingle, G Grant, O S. Harb, M Heiges, F Innamorato, J Iodice, J C. Kissinger, E Kraemer, W Li, J A. Miller, V Nayak, C Pennington, D F. Pinney, D S. Roos, C Ross, C J Stoeckert, Jr,

C Treatman, and H Wang. Plasmodb: a functional genomic database for malaria parasites. *Nucleic Acids Res*, 37(Database issue):D539–D543, Jan 2009. doi: 10.1093/nar/gkn814. URL <http://dx.doi.org/10.1093/nar/gkn814>.

Cecilia Bastos. Small molecule inhibitors of plasmodium falciparum dihydroorotate dehydrogenase, 2011.

Gregory H. Bledsoe. Malaria primer for clinicians in the united states. *South Med J*, 98(12): 1197–204; quiz 1205, 1230, Dec 2005.

J R Bock and Gough D A. Virtual screen for ligands of orphan g protein-coupled receptors. *J Chem Inf Model*, 45(5):1402–1414, 2005. doi: 10.1021/ci050006d. URL <http://dx.doi.org/10.1021/ci050006d>.

H. J. Böhm. On the use of ludi to search the fine chemicals directory for ligands of proteins of known three-dimensional structure. *J Comput Aided Mol Des*, 8(5):623–632, Oct 1994.

Z Bozdech, M Llinés, B L Pulliam, E D Wong, J Zhu, and J L DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS Biol*, 1(1):E5, Oct 2003. doi: 10.1371/journal.pbio.0000005. URL <http://dx.doi.org/10.1371/journal.pbio.0000005>.

D. R. Brooks, P. Wang, M. Read, W. M. Watkins, P. F. Sims, and J. E. Hyde. Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, plasmodium falciparum, with differing resistance to sulfadoxine. *Eur J Biochem*, 224(2):397–405, Sep 1994.

D W A Buchan, A J Shepherd, D Lee, F M G Pearl, S C G Rison, J M Thornton, and C A Orengo. Gene3d: structural assignment for whole genes and genomes using the cath domain structure database. *Genome Res*, 12(3):503–514, Mar 2002. doi: 10.1101/gr.213802. URL <http://dx.doi.org/10.1101/gr.213802>.

P. Bucher, K. Karplus, N. Moeri, and K. Hofmann. A flexible motif search technique based on generalized profiles. *Comput Chem*, 20(1):3–23, Mar 1996.

- D Butina, M D Segall, and K Frankcombe. Predicting adme properties in silico: methods and models. *Drug Discov Today*, 7(11):S83–S88, Jun 2002.
- D. J. Bzik, W. B. Li, T. Horii, and J. Inselburg. Molecular cloning and sequence analysis of the plasmodium falciparum dihydrofolate reductase-thymidylate synthase gene. *Proc Natl Acad Sci U S A*, 84(23):8360–8364, Dec 1987.
- P. R. Caron, M. D. Mullican, R. D. Mashal, K. P. Wilson, M. S. Su, and M. A. Murcko. Chemogenomic approaches to drug discovery. *Curr Opin Chem Biol*, 5(4):464–470, Aug 2001.
- M Cases, R García-Serna, K Hettne, M Weeber, J van der Lei, S Boyer, and J Mestres. Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family. *Curr Top Med Chem*, 5(8):763–772, 2005.
- X. Chen, M. Liu, and M. K. Gilson. Bindingdb: a web-accessible molecular recognition database. *Comb Chem High Throughput Screen*, 4(8):719–725, Dec 2001.
- S Choi, A Pradhan, N L. Hammond, A G. Chittiboyina, B L. Tekwani, and M A. Avery. Design, synthesis, and biological evaluation of plasmodium falciparum lactate dehydrogenase inhibitors. *J Med Chem*, 50(16):3841–3850, Aug 2007. doi: 10.1021/jm070336k. URL <http://dx.doi.org/10.1021/jm070336k>.
- F. Corpet, J. Gouzy, and D. Kahn. The prodrom database of protein domain families. *Nucleic Acids Res*, 26(1):323–326, Jan 1998.
- R D. Cramer, D E. Patterson, and J D. Bunce. Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, 110(18):5959–5967, 1988. doi: 10.1021/ja00226a005. URL <http://pubs.acs.org/doi/abs/10.1021/ja00226a005>.
- J. P. Davis, G. A. Cain, W. J. Pitts, R. L. Magolda, and R. A. Copeland. The immunosuppressive metabolite of leflunomide is a potent inhibitor of human dihydroorotate de-

hydrogenase. *Biochemistry*, 35(4):1270–1273, Jan 1996. doi: 10.1021/bi952168g. URL <http://dx.doi.org/10.1021/bi952168g>.

L. M. Deck, R. E. Royer, B. B. Chamblee, V. M. Hernandez, R. R. Malone, J. E. Torres, L. A. Hunsaker, R. C. Piper, M. T. Makler, and D. L. Vander Jagt. Selective inhibitors of human lactate dehydrogenases and lactate dehydrogenase from the malarial parasite plasmodium falciparum. *J Med Chem*, 41(20):3879–3887, Sep 1998. doi: 10.1021/jm980334n. URL <http://dx.doi.org/10.1021/jm980334n>.

M Dhanawat, N Das, R C Nagarwal, and S. K. Shrivastava. Antimalarial drug development: past to present scenario. *Mini Rev Med Chem*, 9(12):1447–1469, Oct 2009.

K. A. Dill. Additivity principles in biochemistry. *J Biol Chem*, 272(2):701–704, Jan 1997.

S Durdagi, T Mavromoustakos, and M G Papadopoulos. 3d qsar comfa/comsia, molecular docking and molecular dynamics studies of fullerene-based hiv-1 pr inhibitors. *Bioorg Med Chem Lett*, 18(23):6283–6289, Dec 2008. doi: 10.1016/j.bmcl.2008.09.107. URL <http://dx.doi.org/10.1016/j.bmcl.2008.09.107>.

S. R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des*, 11(5):425–445, Sep 1997.

D Erhan, P J L'heureux, S Y Yue, and Y Bengio. Collaborative filtering on a family of biological targets. *J Chem Inf Model*, 46(2):626–635, 2006. doi: 10.1021/ci050367t. URL <http://dx.doi.org/10.1021/ci050367t>.

J. Everse and N. O. Kaplan. Lactate dehydrogenases: structure and function. *Adv Enzymol Relat Areas Mol Biol*, 37:61–133, 1973.

T M Frimurer, T Ulven, C E Elling, L O Gerlach, E Kostenis, and T Högberg. A physico-genetic method to assign ligand-binding relationships between 7tm receptors. *Bioorg Med Chem Lett*, 15(16):3707–3712, Aug 2005. doi: 10.1016/j.bmcl.2005.05.102. URL <http://dx.doi.org/10.1016/j.bmcl.2005.05.102>.

F J Gamo, L M Sanz, Jaume V, C de Cozar, E Alvarez, JL Lavandera, D E Vanderwall, D V S Green, V Kumar, S Hasan, J R Brown, C E Peishoff, L R Cardon, and J F Garcia-Bustos. Thousands of chemical starting points for antimalarial lead identification. *Nature*, 465(7296):305–310, May 2010. doi: 10.1038/nature09107. URL <http://dx.doi.org/10.1038/nature09107>.

M J Gardner, N Hall, E Fung, O White, M Berriman, R W. Hyman, J M. Carlton, A Pain, K E. Nelson, S Bowman, I T. Paulsen, K James, J A. Eisen, K Rutherford, S L. Salzberg, A Craig, S Kyes, M Chan, V Nene, S J. Shallom, B Suh, J Peterson, S Angiuoli, M Pertea, J Allen, J Selengut, D Haft, M W. Mather, A B. Vaidya, D M A. Martin, A H. Fairlamb, M J. Fraunholz, D S. Roos, S A. Ralph, G I. McFadden, L M. Cummings, G M Subramanian, C Mungall, J C Venter, D J. Carucci, S L. Hoffman, C Newbold, R W. Davis, C M. Fraser, and B Barrell. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511, Oct 2002. doi: 10.1038/nature01097. URL <http://dx.doi.org/10.1038/nature01097>.

C. Granchi, S. Bertini, M. Macchia, and F. Minutolo. Inhibitors of lactate dehydrogenase isoforms and their therapeutic potentials. *Curr Med Chem*, 17(7):672–697, 2010.

D. H. Haft, B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White. Tigrfams: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*, 29(1):41–43, Jan 2001.

I Halperin, B Ma, H Wolfson, and R Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, Jun 2002. doi: 10.1002/prot.10115. URL <http://dx.doi.org/10.1002/prot.10115>.

T. Hänscheid. Diagnosis of malaria: a review of alternatives to conventional microscopy. *Clin Lab Haematol*, 21(4):235–245, Aug 1999.

C Hoppe, C Steinbeck, and G Wohlfahrt. Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials. *J Mol Graph Model*, 24(5):328–340, Mar 2006. doi: 10.1016/j.jmghm.2005.09.013. URL <http://dx.doi.org/10.1016/j.jmghm.2005.09.013>.

S Hunter, P Jones, A Mitchell, R Apweiler, T K Attwood, A Bateman, T Bernard, D Binns, P Bork, S Burge, E de Castro, P Coggill, M Corbett, U Das, L Daugherty, L Duquenne, R D Finn, M Fraser, J Gough, D Haft, N Hulo, D Kahn, E Kelly, I Letunic, D Lonsdale, R Lopez, M Madera, J Maslen, C McAnulla, J McDowall, C McMenamin, H Mi, P Mutowo-Muellenet, N Mulder, D Natale, C Orengo, S Pesseat, M Punta, A F Quinn, C Rivoire, A Sangrador-Vegas, J D Selengut, C J A Sigrist, M Scheremetjew, J Tate, M Thimmajananathan, P D Thomas, C H Wu, C Yeats, and SY Yong. Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*, 40(Database issue):D306–D312, Jan 2012. doi: 10.1093/nar/gkr948. URL <http://dx.doi.org/10.1093/nar/gkr948>.

J A Jadwin, M Ogiue-Ikeda, and K Machida. The application of modular protein domains in proteomics. *FEBS Lett*, 586(17):2586–2596, Aug 2012. doi: 10.1016/j.febslet.2012.04.019. URL <http://dx.doi.org/10.1016/j.febslet.2012.04.019>.

D. L. Vander Jagt, L. A. Hunsaker, and J. E. Heidrich. Partial purification and characterization of lactate dehydrogenase from plasmodium falciparum. *Mol Biochem Parasitol*, 4(5-6):255–264, Dec 1981.

G. Jones, P. Willett, and R. C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol*, 245(1):43–53, Jan 1995.

M. E. Jones. Pyrimidine nucleotide biosynthesis in animals: genes, enzymes, and regulation of ump biosynthesis. *Annu Rev Biochem*, 49:

253–279, 1980. doi: 10.1146/annurev.bi.49.070180.001345. URL
<http://dx.doi.org/10.1146/annurev.bi.49.070180.001345>.

W. L. Jorgensen. Rusting of the lock and key model for protein-ligand binding. *Science*, 254 (5034):954–955, Nov 1991.

F Joubert, C M Harrison, R J Koegelenberg, C J Odendaal, and T A P de Beer. Discovery: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. *Malar J*, 8:178, 2009. doi: 10.1186/1475-2875-8-178. URL
<http://dx.doi.org/10.1186/1475-2875-8-178>.

C Francek Z Chen J Buenviaje D Plouffe E Winzeler A Brinker T Diagana J Taylor R Glynne A Chatterjee K Kuhlen K Gagaring, R Borboa. Novartis-gnf malaria box. Genomics Institute of the Novartis Research Foundation (GNF), 10675 John Jay Hopkins Drive, San Diego CA 92121, USA and Novartis Institute for Tropical Disease, 10 Biopolis Road, Chromos , 05-01, 138 670 Singapore.

S H I Kappe, A M Vaughan, J A Boddey, and A F Cowman. That was then but this is now: malaria research in the time of an eradication agenda. *Science*, 328(5980):862–866, May 2010. doi: 10.1126/science.1184785. URL
<http://dx.doi.org/10.1126/science.1184785>.

D E. Koshland. The key-lock theory and the induced fit theory. *Angewandte Chemie International Edition in English*, 33(23-24):2375–2378, 1995. ISSN 1521-3773. doi: 10.1002/anie.199423751. URL <http://dx.doi.org/10.1002/anie.199423751>.

A Koutsoukas, B Simms, J Kirchmair, P J Bond, A V Whitmore, S Zimmer, M P Young, J L Jenkins, M Glick, R C Glen, and A Bender. From in silico target prediction to multi-target drug design: Current databases, methods and applications. *J Proteomics*, May 2011. doi: 10.1016/j.jprot.2011.05.011. URL <http://dx.doi.org/10.1016/j.jprot.2011.05.011>.

R T Kroemer. Structure-based drug design: docking and scoring. *Curr Protein Pept Sci*, 8 (4):312–328, Aug 2007.

D. J. Krogstad, I. Y. Gluzman, D. E. Kyle, A. M. Oduola, S. K. Martin, W. K. Milhous, and P. H. Schlesinger. Efflux of chloroquine from plasmodium falciparum: mechanism of chloroquine resistance. *Science*, 238(4831):1283–1285, Nov 1987.

S. T. Agnandji B. Lell, S S Soulanoudjingar, J F Fernandes, B P Abossolo, C Conzelmann, B G N O Methogo, Y Doucka, A Flamen, B Mordmüller, S Issifou, P G Kremsner, J Sacarlal, P Aide, M Lanaspá, J J Aponte, A Nhamuave, D Quelhas, Q Bassat, S Mandjate, E Macete, P Alonso, S Abdulla, N Salim, O Juma, M Shomari, K Shubis, F Machera, A S Hamad, R Minja, A Mtoro, A Sykes, S Ahmed, A M Urassa, A M Ali, G Mwangoka, M Tanner, H Tinto, U D’Alessandro, H Sorgho, I Valea, M C Tahita, W Kaboré, S Ouédraogo, Y Sandrine, R T Guiguemdé, J B Ouédraogo, M J Hamel, S Kariuki, C Odero, M Oneko, K Otieno, N Awino, J Omoto, J Williamson, V Muturi-Kioi, K F Laserson, L Slutsker, W Otieno, L Otieno, O Nekoye, S Gondi, A Otieno, B Ogutu, R Wasuna, V Owira, D Jones, A A Onyango, P Njuguna, R Chilengi, P Akoo, C Kerubo, J Gitaka, C Maingi, T Lang, A Olotu, B Tsofa, P Bejon, N Peshu, K Marsh, S Owusu-Agyei, K P Asante, K Osei-Kwakye, O Boahen, S Ayamba, K Kayan, R Owusu-Ofori, D Dosoo, I Asante, G Adjei, G Adjei, D Chandramohan, B Greenwood, J Lusingu, S Gesase, A Malabeja, O Abdul, H Kilavo, C Mahende, E Liheluka, M Lemnge, T Theander, C Drakeley, D Ansong, T Agbenyega, S Adjei, H O Boateng, T Rettig, J Bawa, J Sylverken, D Sambian, A Agyekum, L Owusu, F Martinson, I Hoffman, T Mvalo, P Kamthunzi, R Nkomo, A Msika, A Jumbe, N Chome, D Nyakuipa, J Chintedza, W. R Ballou, M Bruls, J Cohen, Y Guerra, E Jongert, D Lapierre, A Leach, M Lievens, O Ofori-Anyinam, J Vekemans, T Carter, D Leboulleux, C Loucq, A Radford, B Savarese, D Schellenberg, M Sillman, P Vansadia, and S. Clinical Trials Partnership R. T. S. First results of phase 3 trial of rts,s/as01 malaria vaccine in african children. *N Engl J Med*, 365(20):1863–1875, Nov 2011.

M L. Lolli, M Giorgis, P Tosco, A Foti, R Fruttero, and A Gasco. New inhibitors of dihydroorotate dehydrogenase (dhodh) based on the 4-hydroxy-1,2,5-oxadiazol-3-yl (hydroxyfurazanyl) scaffold. *Eur J Med*

Chem, 49:102–109, Mar 2012. doi: 10.1016/j.ejmech.2011.12.038. URL <http://dx.doi.org/10.1016/j.ejmech.2011.12.038>.

María P. Magariños, Santiago J. Carmona, Gregory J. Crowther, Stuart A. Ralph, David S. Roos, Dhanasekaran Shanmugam, Wesley C. Van Voorhis, and Fernán Agüero. Tdr targets: a chemogenomics resource for neglected diseases. *Nucleic Acids Res*, 40(Database issue):D1118–D1127, Jan 2012. doi: 10.1093/nar/gkr1053. URL <http://dx.doi.org/10.1093/nar/gkr1053>.

M. T. Makler, J. M. Ries, J. A. Williams, J. E. Bancroft, R. C. Piper, B. L. Gibbins, and D. J. Hinrichs. Parasite lactate dehydrogenase as an assay for plasmodium falciparum drug sensitivity. *Am J Trop Med Hyg*, 48(6):739–741, Jun 1993.

C. L. Markert. Lactate dehydrogenase. biochemistry and function of lactate dehydrogenase. *Cell Biochem Funct*, 2(3):131–134, Jul 1984. doi: 10.1002/cbf.290020302. URL <http://dx.doi.org/10.1002/cbf.290020302>.

H Mi, Q Dong, A Muruganujan, P Gaudet, S Lewis, and P D Thomas. Panther version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium. *Nucleic Acids Res*, 38(Database issue):D204–D210, Jan 2010. doi: 10.1093/nar/gkp1019. URL <http://dx.doi.org/10.1093/nar/gkp1019>.

L. H. Miller, M. F. Good, and G. Milon. Malaria pathogenesis. *Science*, 264(5167):1878–1883, Jun 1994.

T Naumann and H Matter. Structural classification of protein kinases using 3d molecular interaction field analysis of their ligand binding sites: target family landscapes. *J Med Chem*, 45(12):2366–2378, Jun 2002.

Nidhi, M Glick, J W Davies, and J L Jenkins. Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. *J Chem Inf Model*, 46(3):1124–1133, 2006. doi: 10.1021/ci060003g. URL <http://dx.doi.org/10.1021/ci060003g>.

World Health Organization(WHO). World malaria report, 2012. URL <http://www.who.int/malaria/publications/en/>.

V Patel, M Booker, M Kramer, L Ross, C A Celatka, L M Kennedy, J D Dvorin, M T Duraisingh, P Sliz, D F Wirth, and J Clardy. Identification and characterization of small molecule inhibitors of plasmodium falciparum dihydroorotate dehydrogenase. *J Biol Chem*, 283(50):35078–35085, Dec 2008. doi: 10.1074/jbc.M804990200. URL <http://dx.doi.org/10.1074/jbc.M804990200>.

P.Baldi and R.Nasr. When is chemical similarity significant? the statistical distribution of chemical similarity scores and its extreme values. *J Chem Inf Model*, 50(7):1205–1222, Jul 2010. doi: 10.1021/ci100010v. URL <http://dx.doi.org/10.1021/ci100010v>.

E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. Interproscan: protein domains identifier. *Nucleic Acids Res*, 33 (Web Server issue):W116–W120, Jul 2005. doi: 10.1093/nar/gki442. URL <http://dx.doi.org/10.1093/nar/gki442>.

P. Reyes, P. K. Rathod, D. J. Sanchez, J. E. Mrema, K. H. Rieckmann, and H. G. Heidrich. Enzymes of purine and pyrimidine metabolism from the human malaria parasite, plasmodium falciparum. *Mol Biochem Parasitol*, 5(5):275–290, May 1982.

R G Ridley. Medical need, scientific opportunity and the drive for antimalarial drugs. *Nature*, 415(6872):686–693, Feb 2002. doi: 10.1038/415686a. URL <http://dx.doi.org/10.1038/415686a>.

D. Rognan. Chemogenomic approaches to rational drug design. *Br J Pharmacol*, 152(1):38–52, Sep 2007. doi: 10.1038/sj.bjp.0707307. URL <http://dx.doi.org/10.1038/sj.bjp.0707307>.

J. Sadowski and H. Kubinyi. A scoring scheme for discriminating between drugs and nondrugs. *J Med Chem*, 41(18):3325–3329, Aug 1998. doi: 10.1021/jm9706776. URL <http://dx.doi.org/10.1021/jm9706776>.

- J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting. Smart, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*, 95(11):5857–5864, May 1998.
- L Schwartz, G V Brown, B Genton, and V S Moorthy. A review of malaria vaccine clinical projects based on the who rainbow table. *Malar J*, 11(1):11, Jan 2012. doi: 10.1186/1475-2875-11-11. URL <http://dx.doi.org/10.1186/1475-2875-11-11>.
- P. Scordis, D. R. Flower, and T. K. Attwood. Fingerprints can: intelligent searching of the prints motif database. *Bioinformatics*, 15(10):799–806, Oct 1999.
- M. S. Smyth and J. H. Martin. x ray crystallography. *Mol Pathol*, 53(1):8–14, Feb 2000.
- E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Res*, 26(1):320–322, Jan 1998.
- T M Steindl, D Schuster, C Laggner, and T Langer. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J Chem Inf Model*, 46(5):2146–2157, 2006. doi: 10.1021/ci6002043. URL <http://dx.doi.org/10.1021/ci6002043>.
- S. A. Sundberg. High-throughput and ultra-high-throughput screening: solution- and cell-based approaches. *Curr Opin Biotechnol*, 11(1):47–53, Feb 2000.
- J S Surgand, J Rodrigo, E Kellenberger, and D Rognan. A chemogenomic analysis of the transmembrane binding cavity of human g-protein-coupled receptors. *Proteins*, 62(2):509–538, Feb 2006. doi: 10.1002/prot.20768. URL <http://dx.doi.org/10.1002/prot.20768>.
- S J Teague. Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov*, 2(7):527–541, Jul 2003. doi: 10.1038/nrd1129. URL <http://dx.doi.org/10.1038/nrd1129>.
- A Teklehaimanot and P Mejia. Malaria and poverty. *Ann N Y Acad Sci*, 1136:32–37, 2008. doi: 10.1196/annals.1425.037. URL <http://dx.doi.org/10.1196/annals.1425.037>.
- Siva Sanjeeva Rao Thunuguntla. Dihydroorotate dehydrogenase inhibitors, 2010.

- T. Triglia and A. F. Cowman. Primary structure and expression of the dihydropteroate synthetase gene of plasmodium falciparum. *Proc Natl Acad Sci U S A*, 91(15):7149–7153, Jul 1994.
- C. Y. Wan and T. A. Wilkins. Spermidine facilitates pcr amplification of target dna. *PCR Methods Appl*, 3(3):208–210, Dec 1993.
- Y Wang, J Xiao, T O. Suzek, J Zhang, J Wang, and S H. Bryant. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*, 37(Web Server issue):W623–W633, Jul 2009. doi: 10.1093/nar/gkp456. URL <http://dx.doi.org/10.1093/nar/gkp456>.
- Y Wang, J Xiao, T O. Suzek, J Zhang, J Wang, Z Zhou, L Han, K Karapetyan, S Dracheva, B A. Shoemaker, E Bolton, A Gindulyte, and S H. Bryant. Pubchem’s bioassay database. *Nucleic Acids Res*, 40(Database issue):D400–D412, Jan 2012. doi: 10.1093/nar/gkr1132. URL <http://dx.doi.org/10.1093/nar/gkr1132>.
- N J White. Antimalarial drug resistance. *J Clin Invest*, 113(8):1084–1092, Apr 2004. doi: 10.1172/JCI21682. URL <http://dx.doi.org/10.1172/JCI21682>.
- P Willett. Similarity-based virtual screening using 2d fingerprints. *Drug Discov Today*, 11(23-24):1046–1053, Dec 2006. doi: 10.1016/j.drudis.2006.10.005. URL <http://dx.doi.org/10.1016/j.drudis.2006.10.005>.
- Kurt Wuthrich. *NMR of proteins and nucleic acids*. Wiley, 1986.
- X Xia, E G Maliski, P Gallant, and D Rogers. Classification of kinase inhibitors using a bayesian model. *J Med Chem*, 47(18):4463–4470, Aug 2004. doi: 10.1021/jm0303195. URL <http://dx.doi.org/10.1021/jm0303195>.