In Silico Resources for Malaria Drug Discovery

Pieter B. Burger and Fourie Joubert
University of Pretoria
South Africa

1. Introduction

Drugs currently in use against the malaria parasite have been derived from known natural compounds, or discovered serendipitously. The completion of the *Plasmodium falciparum* genome project at the turn of the century (Gardner *et al.*, 2002) raised great hopes for the identification of a wealth of targets against which to design new and novel drugs. However, it soon became apparent that the challenges associated with the annotation of the malaria proteome was creating a significant hurdle to be overcome already in the early stages. In addition, the complexity introduced by the multitude of factors that need consideration in selecting suitable drug targets made it difficult to perform the selection of target proteins in a high-throughput automated fashion. The need for the development of effective *in silico* approaches required to successfully address these problems were identified and addressed by the malaria community.

This chapter focuses on *in silico* resources that are available to support researchers working in the area of malaria drug discovery. It aims to facilitate the researcher's entry into early phase drug discovery by lowering the initial barrier to data mining sometimes perceived as a daunting task due to information overload. While *in silico* experimentation is not intended to replace detailed experimental work, it may be extremely useful in decreasing the size of the protein and chemical space to be investigated *in vitro*, and may help guide the experimentalist's decision making during the selection process. These resources are intended to assist the researcher in rationally selecting a relatively small number of targets out of the whole for further detailed investigation. The *in silico* resources discussed here vary from primary data repositories, to advanced data mining systems, and provide information on a variety of different levels in discovery research.

Genome databases are often the first point of access. Whereas they always contain the primary data generated during the genome project, they usually contain significant amounts of annotations regarding gene structures and gene products including nucleotide and protein sequences. The annotation detail may vary, as is the case with the malaria genomes. *Plasmodium falciparum* has been annotated extensively, but the level of annotation for the other Plasmodia may range widely. The genome database for malaria parasites additionally provides extensive information on gene variation, protein features, expression and a wide variety of other molecular properties and is a rich resource from which to embark on further study. Each entry also provides many external links for exploration.

The more specialized databases are typically derived from the data present in the genome database, and provide focussed information on specific aspects related to the molecules available. There are several resources containing information around metabolic pathways of the malaria parasites, assisting researchers in understanding the pathways present in the organisms, and the role of specific proteins or compounds in the organisms' metabolic activity, which is especially useful in deciding on aspects of the parasite metabolism that may be selected for interference. Protein-protein interaction resources highlight interacting proteins, but additionally allow the researcher to begin understanding the interaction and regulatory network that the parasite requires for growth and survival, together with possible ways to manipulate this homeostasis. Gene expression resources provide insights into the regulatory mechanisms of gene expression, especially in terms of changes in expression profiles over the different stages of the parasite. Functional prediction resources attempt to expand the annotation of novel and unassigned proteins produced by the parasites, particularly by attempting to use methods that may be independent from the protein sequence of the molecule using eg. guilt-by-association approaches. Protein structure databases provide the researcher with experimentally-determined structures or models to help understand molecular mechanisms, and for further use in comparative undertakings and molecular modeling projects. Literature resources focus on articles published on the many aspects of malaria and its molecules, and may attempt to automatically extract value-added information from the text. Drug discovery resources attempt to integrate many of the aspects already mentioned here into a single system, where researchers may perform the selection and scoring of possible drug targets and lead compounds in an automated or semi-automated fashion.

A summary of the resource discussed in this Chapter is presented in Table 1.

Name	URL
PlasmoDB	http://www.plasmodb.org
Malaria Parasite Metabolic Pathways	http://sites.huji.ac.il/malaria
PlasmoCyc	http://plasmocyc.stanford.edu
KEGG	http://www.genome.jp/kegg
PlasmoMap	http://www.cbil.upenn.edu/plasmoMAP
IDC Strain Comparison Database	http://malaria.ucsf.edu
PlasmoDraft	http://www.atgc-montpellier.fr/PlasmoDraft
Protein Structure Database	http://www.pdb.org
TDI Kernel	http://tropicaldisease.org/kernel
ModBase	http://modbase.compbio.ucsf.edu
Malaria Literature Database	http://carrier.gnf.org/publications/Py
ChEMBL	http://www.ebi.ac.uk/chembldb
TDR Targets	http://tdrtargets.org
Discovery	http://discovery.bi.up.ac.za

Table 1. A summary of resources discussed in this chapter

2. Malaria genome databases

2.1 PlasmoDB

URL: http://www.plasmodb.org

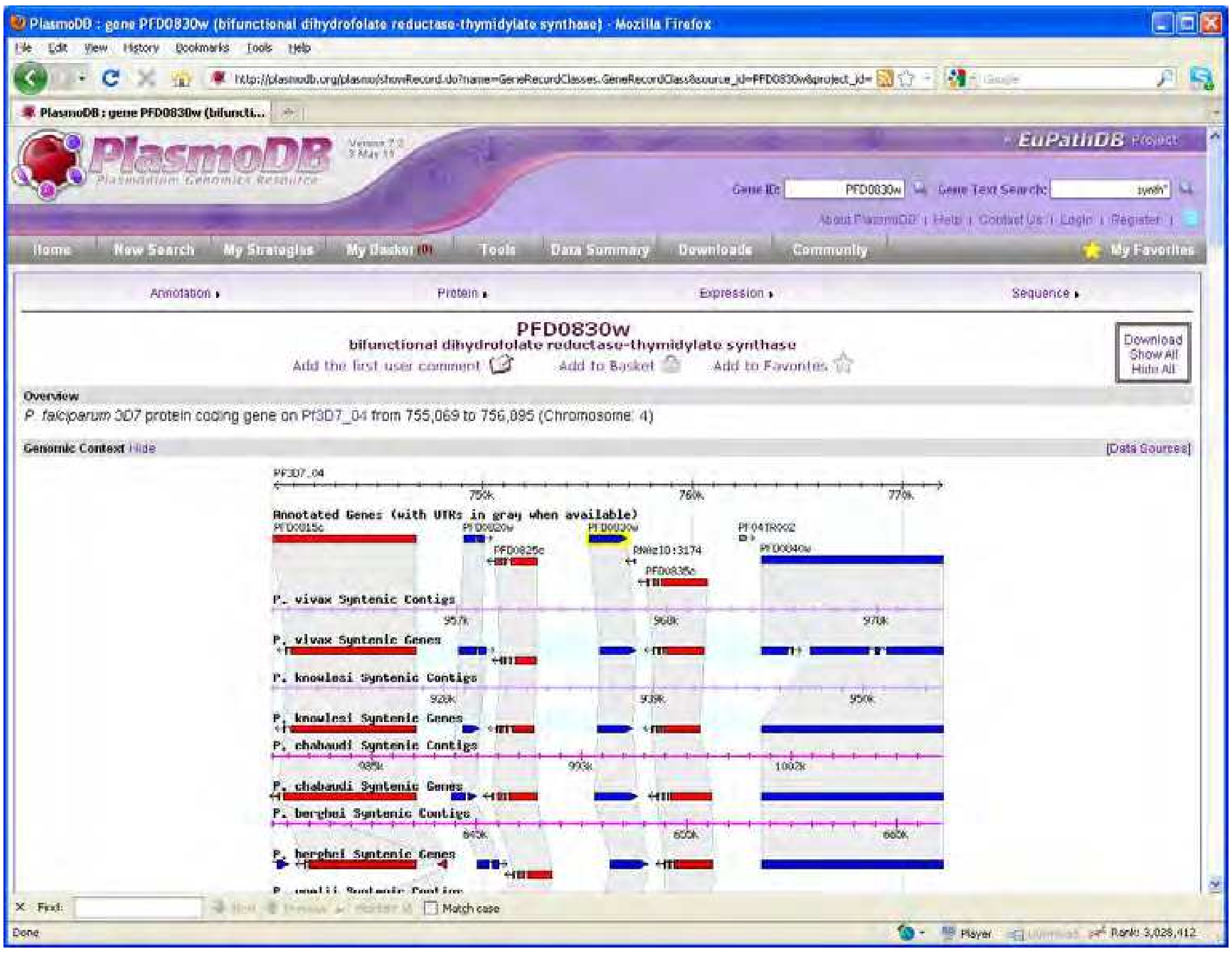
PlasmoDB (Aurrecoechea *et al.*, 2009) is the primary resource for malaria genome information, and is part of the larger EuPathDB project (Aurrecoechea *et al.*, 2010) which is focussed on eukaryotic pathogens. The PlasmoDB resource contains genome data for *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi* and *P. knowlesi*. PlasmoDB is regarded as the primary distribution point for malaria sequence data, and while genome and proteome data are available for download, an advanced data mining system is provided for complex queries in gene, protein and related features. The most basic mechanism to access the PlasmoDB site is by searching using either a known accession number or a keyword. The more useful approach is to perform a text-based query, where series of species and data fields may be selected for searching. This will provide the user with a list of entries matching the search criteria. When an entry is selected, a page for the molecule is displayed, containing the genomic context, together with annotation, protein, expression and sequence information.

Each of these categories has a rich subset of information, which may be selected for display, including single nucleotide polymorphisms, sequence aligment vs. the other malaria species, links to annotation information in other databases, orthologs and paralogs, ontology information, metabolism information, physical-chemical information, structural information, immunological information, microarray and proteomics results (Figure 1). Some of the additional data is generated by the PlasmoDB group, while other data is linked to 3rd-party sites and research groups.

While a simple text search may be useful in straight-forward cases, the powerful query engine underlying the resource provides the user with the capability to select molecules based on a range of properties, in a sequential or branched fashion. Filters may be based on a wide range of properties as described in the previous paragraph. As an example, a scenario is illustrated where the user is interested in finding all *P. falciparum* proteins matching the text term "synth*" where the effect of SNPs may be predicted on a 3D level. An initial query may be performed by selecting all proteins containing the annotation "synth*". The user is then presented with a graphical view of his query results, showing the number of initial results (119 in this case). An additional intersect filter may then be added to only show proteins where SNP information is available (68 genes). This may then be further filtered to only show proteins where predicted 3D-structures are also available (12 genes). The complete search strategy is visualized graphically as a pipeline (Figure 2), where the user interactively may alter points in the search strategy, create branches and set weighting for different aspects of the search. In a search strategy, unions and intersects will sum the weights, giving higher scores to items found in multiple searches. This interface is regarded as one of the most powerful, user-friendly and intuitive for all biological databases.

Users may create a personal account on the site, and search strategies as well as results may be saved in this way. In addition to the database searches, a series of tools are also available at PlasmoDB. These tools enable the user to perform BLAST searches against the malaria data, retrieve sequences by lists of gene IDs, access links to the PubMed and Entrez sites, access the malaria genome data in the GBrowse genome browser interface (Stein *et al.*, 2002), predict apicoplast-targeting signals, predict mitochondrial transit peptides and access the PlasmoCyc metabolic pathway annotations. Additionally, a web services WADL (Web Application Description Language) specification for searches against PlasmoDB is provided.

PlasmoDB is carefully curated, highly quality controlled and updated at a very regular frequency. Information about the available data is easily accessed from the data summary section on the front page of the resource. PlasmoDB would be the first resource to visit for a researcher embarking on a malaria drug discovery project, to obtain data about the availability of genes and proteins of interest, together with the relevant annotations and value-added information.



PlasmoDB : gene PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase) - Mozilla Firefox

http://plasmodb.org/plasmo/showRecord.do?name=GeneRecordClasses.GeneRecordClass&source_id=PFD0830w&project_id=

Gene ID: PFD0830w

Gene Text Search: synth*

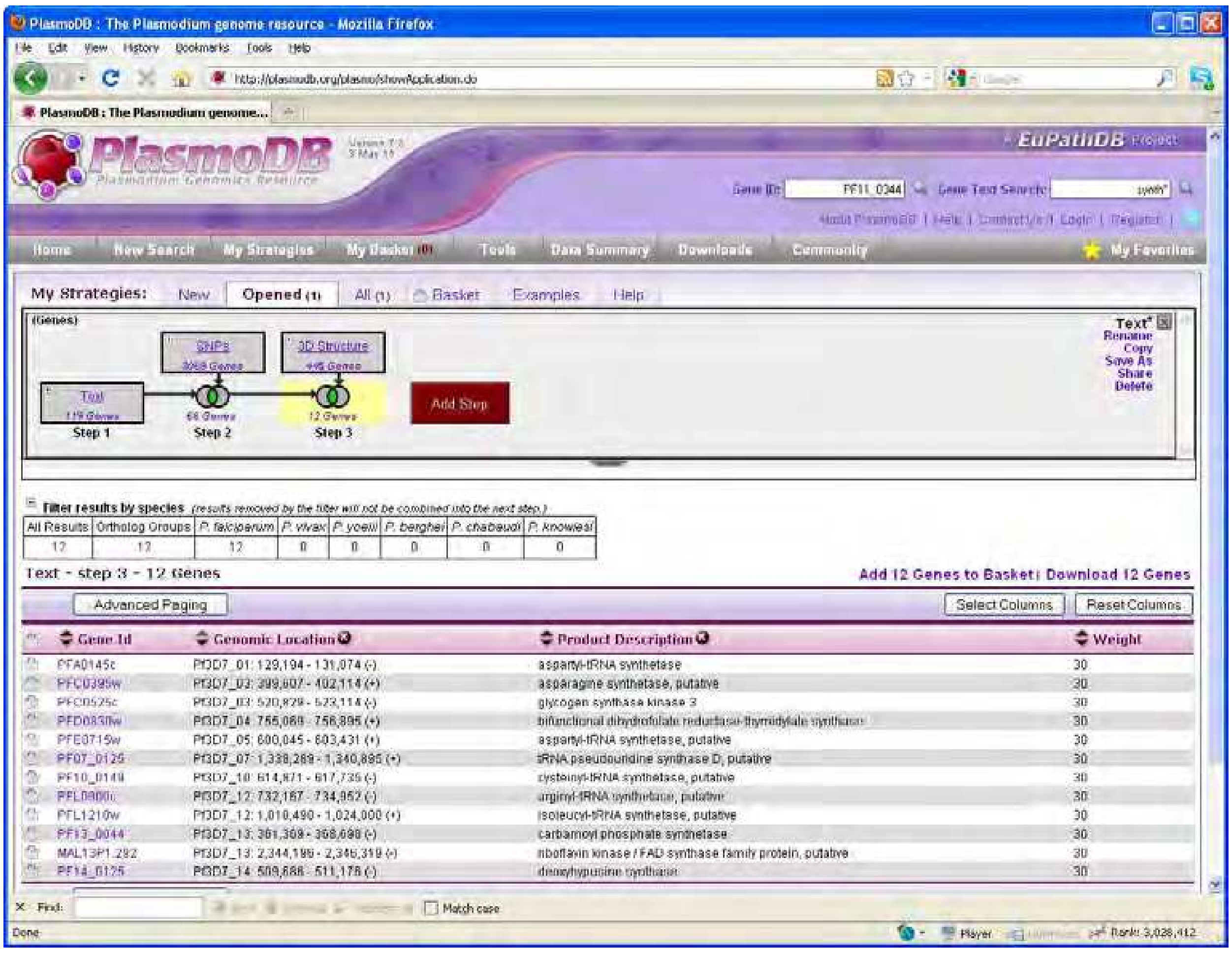
PFD0830w

bifunctional dihydrofolate reductase-thymidylate synthase

Overview

P. falciparum 3D7 protein coding gene on Pf3D7_04 from 755,069 to 756,095 (Chromosome: 4)

Fig. 1. The result view for the protein PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase from *P. falciparum*) in the PlasmoDB, showing the genomic context together with syntenic genes in other species. The rest of the detailed gene information mentioned in the text is not visible in this figure but continues lower down the browser screen.



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

All Results	Ortholog Groups	*P. falciparum*	*P. vivax*	*P. yoelii*	*P. berghei*	*P. chabaudi*	*P. knowlesi*
12	12	12	0	0	0	0	0

Text - step 3 - 12 Genes

Add 12 Genes to Basket | Download 12 Genes

Gene Id	Genomic Location	Product Description	Weight
PFA0145c	Pf3D7_01: 129,194 - 131,074 (-)	aspartyl-tRNA synthetase	30
PFC0395w	Pf3D7_03: 399,607 - 402,114 (+)	asparagine synthetase, putative	30
PFC0525c	Pf3D7_03: 520,829 - 523,114 (-)	glycogen synthase kinase 3	30
PFD0830w	Pf3D7_04: 755,060 - 756,895 (+)	bifunctional dihydrofolate reductase-thymidylate synthase	30
PFE0715w	Pf3D7_05: 600,045 - 603,431 (+)	aspartyl-tRNA synthetase, putative	30
PF07_0125	Pf3D7_07: 1,338,289 - 1,340,895 (+)	tRNA pseudouridine synthase D, putative	30
PF10_0149	Pf3D7_10: 614,871 - 617,735 (-)	cysteinyl-tRNA synthetase, putative	30
PFL0900c	Pf3D7_12: 732,167 - 734,952 (-)	arginyl-tRNA synthetase, putative	30
PFL1210w	Pf3D7_12: 1,010,490 - 1,024,000 (+)	isoleucyl-tRNA synthetase, putative	30
PF13_0044	Pf3D7_13: 361,309 - 368,698 (-)	carbamoyl phosphate synthetase	30
MAL13P1.282	Pf3D7_13: 2,344,186 - 2,346,319 (-)	riboflavin kinase / FAD synthase family protein, putative	30
PF14_0125	Pf3D7_14: 509,688 - 511,178 (-)	deoxyhypusine synthase	30

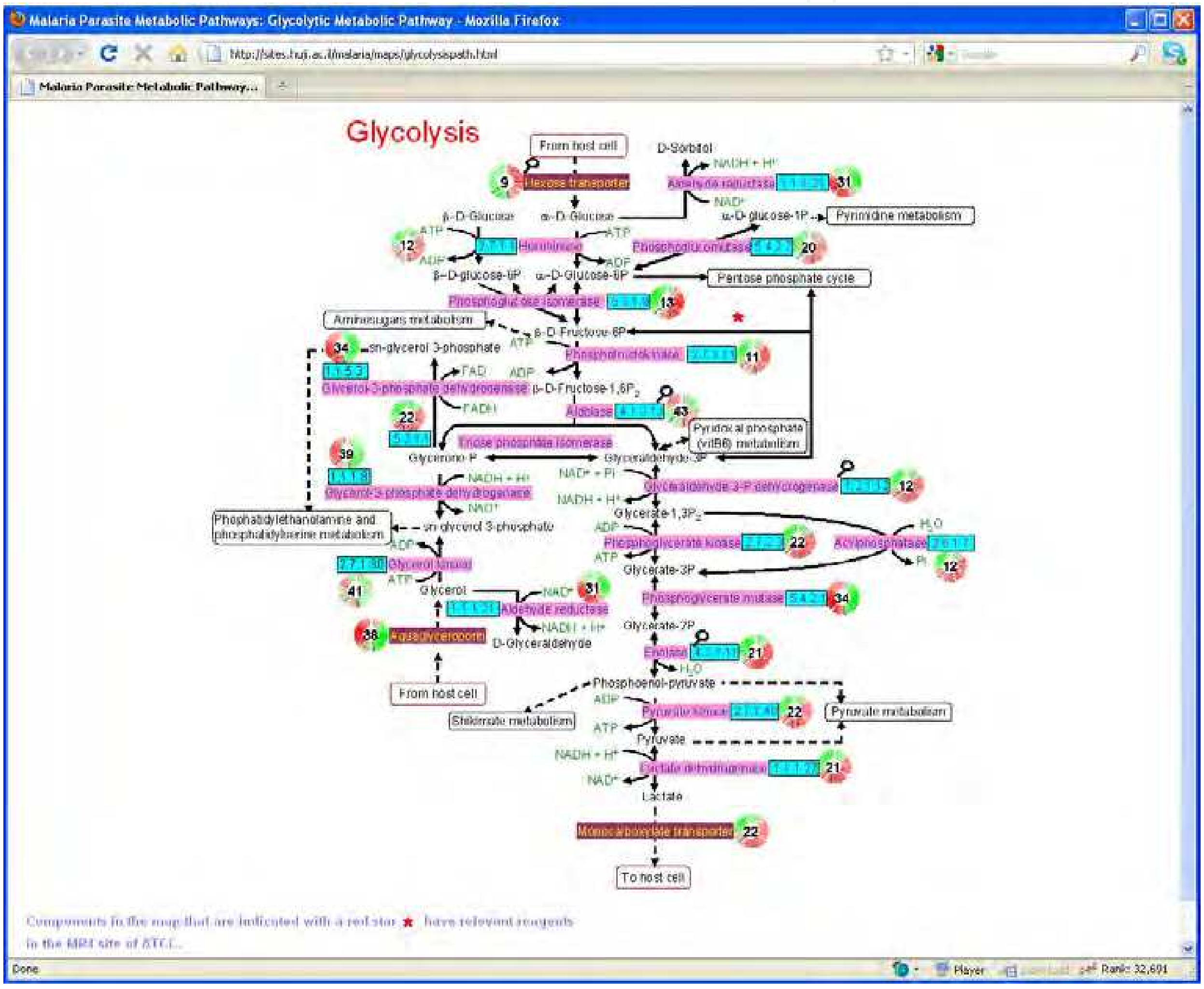
Fig. 2. The advanced result display interface of PlasmoDB, at the top graphically showing the data query pipeline with filtering steps performed using an initial text query followed by the availability of SNP information and 3D structural information. The resulting hits from the query are shown at the bottom. These steps may be edited again, or further branched and expanded.

3. Malaria metabolism databases

3.1 Malaria Parasite Metabolic Pathways (MPMP)

URL: http://sites.huji.ac.il/malaria

The Malaria Parasite Metabolic Pathways resource is maintained by Hagai Ginsburg, and is the most complete and up-to-date source of information for metabolic pathways in malaria (Ginsburg, 2009). Researchers may be interested in exploring information about the different enzymes in a pathway together with their characteristics, or alternatively in seeing the role that a specific enzyme plays in one or more pathways. MPMP has been constructed based on enzymes annotated in the parasite and on pathways known to occur in unicellular eukaryotes. While the pathways have been cross-checked against PlasmoDB, not all enzymes and reactions have been validated in malaria. The site also includes information related to cell-cell interactions, invasion of the erythrocyte and transport functions. The site provides information in a hierarchical fashion, starting with grouped pathways based on chemical component or biological process. This is followed by specific pathways or process, chemical structures and enzymes together with their genes. To browse the site, the user selects a primary category, such as carbohydrates. This is followed by the selection of a specific category. In this example, first carbohydrate metabolism and then specifically the glycolysis pathway was selected for visualization. A graphical overview of the glycolysis pathway is displayed. The overview shows the enzyme names, EC numbers, co-factors and metabolites of the pathway, as well as links to other metabolic pathways (Figure 3). The site may also be searched using enzyme names, EC numbers, other protein names or the names of metabolites.



Components in the map that are indicated with a red star ★ have relevant reagents in the MR4 site of ATCC.

Fig. 3. The glycolysis pathway, as shown in Malaria Parasite Metabolic Pathways. The enzymes and well as chemical compounds and co-factors are visible. A small glyph showing stage-specific expression information is also shown where available.

From each enzyme, links are available to a set of other metabolic pathways databases including BRENDA (http://www.brenda.uni-koeln.de), ExPASy ENZYME (http://www.expasy.org/enzyme) and the IUBMB reaction schemes (http://www.chem.qmul.ac.uk/iubmb/enzyme/reaction). Links are also provided to PlasmoDB (http://plasmodb.org), and *P. falciparum* GeneDB (http://www.genedb.org). From metabolites, links are provided to KEGG (http://www.genome.jp/kegg) which provides chemical structures and formulas. The links provide access to a wealth of information including enzymatic activity, enzyme assays, kinetic parameters and inhibitors. Stage-dependent trancription of each enzyme is also shown as a pie-chart, based on the DeRisi/UCSF transcriptome database (http://malaria.ucsf.edu).

3.2 PlasmoCyc

URL: http://plasmocyc.stanford.edu

PlasmoCyc (Yeh *et al.*, 2004) forms part of the larger MetaCyc resource, where metabolic pathways for a range of organisms have been modeled. PlasmoCyc was initially constructed automatically using annotated EC terms as input to the PathoLogic tools (Caspi *et al.*, 2006). Additional enzymes and reactions were then manually added afterwards based on information obtained from literature. The pathways constructed in this fashion were then further expanded using data from the Malaria Parasite Metabolic Pathways resource. The PlasmoCyc site additionally hosts a set of potentially-interesting drug target proteins, together with motivations for their choice and literature references. PlasmoCyc may be accessed by searching for an enzyme name or EC number, using an ontology browser, or choosing from a list of pathways, proteins or compounds. Pathways may be viewed at different levels of detail. The initial view shows enzyme names, and the next level adds PlasmoDB IDs, EC numbers, compounds and co-factors. The subsequent level provides chemical structures of the compounds. In this example, the glycolysis pathway was selected for display (Figure 4). A useful functionality is the species comparison tool, where metabolic pathways may be compared between several different species to test for differences in the presence of pathway components and to identify unique enzymes for target selection.

3.3 Kyoto Encyclopedia of Genes and Genomes (KEGG)

URL: http://www.genome.jp/kegg

KEGG is a very comprehensive database on metabolism in many species, and includes malaria. KEGG includes information for pathways, diseases, drugs, orthology, genes, genomes, compounds and reactions (Kanehisa & Goto, 2000; Kanehisa *et al.*, 2010; Kanehisa *et al.*, 2006). KEGG contains pathway representations for *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi* and *P. knowlesi*. KEGG may be accessed using various approaches. A common approach for browsing pathways is to select the KEGG pathway entry point, followed by a reference pathway of interest. Subsequently, an organism may be selected. Enzymes identified in the organism are then highlighted using color coding. In this example, the glycolysis pathway was selected for display (Figure 5). Selecting an enzyme on the pathway will provide a detailed page of information on the enzyme, including the sequence, known motifs and links to other databases such as SSDB, GenBank, PlasmoDB, GeneDB and UniProt. Selecting a compound will provide a detailed page including physical-chemical properties, the structure, reactions linked to the compound, pathways linked to the compound, enzymes linked to the compound and links to other databases such as CAS (http://www.cas.org), PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChEBI (http://www.ebi.ac.uk/chebi/), KNApSAcK (http://kanaya.naist.jp/KNApSAcK/) and PDB (http://www.pdb.org).

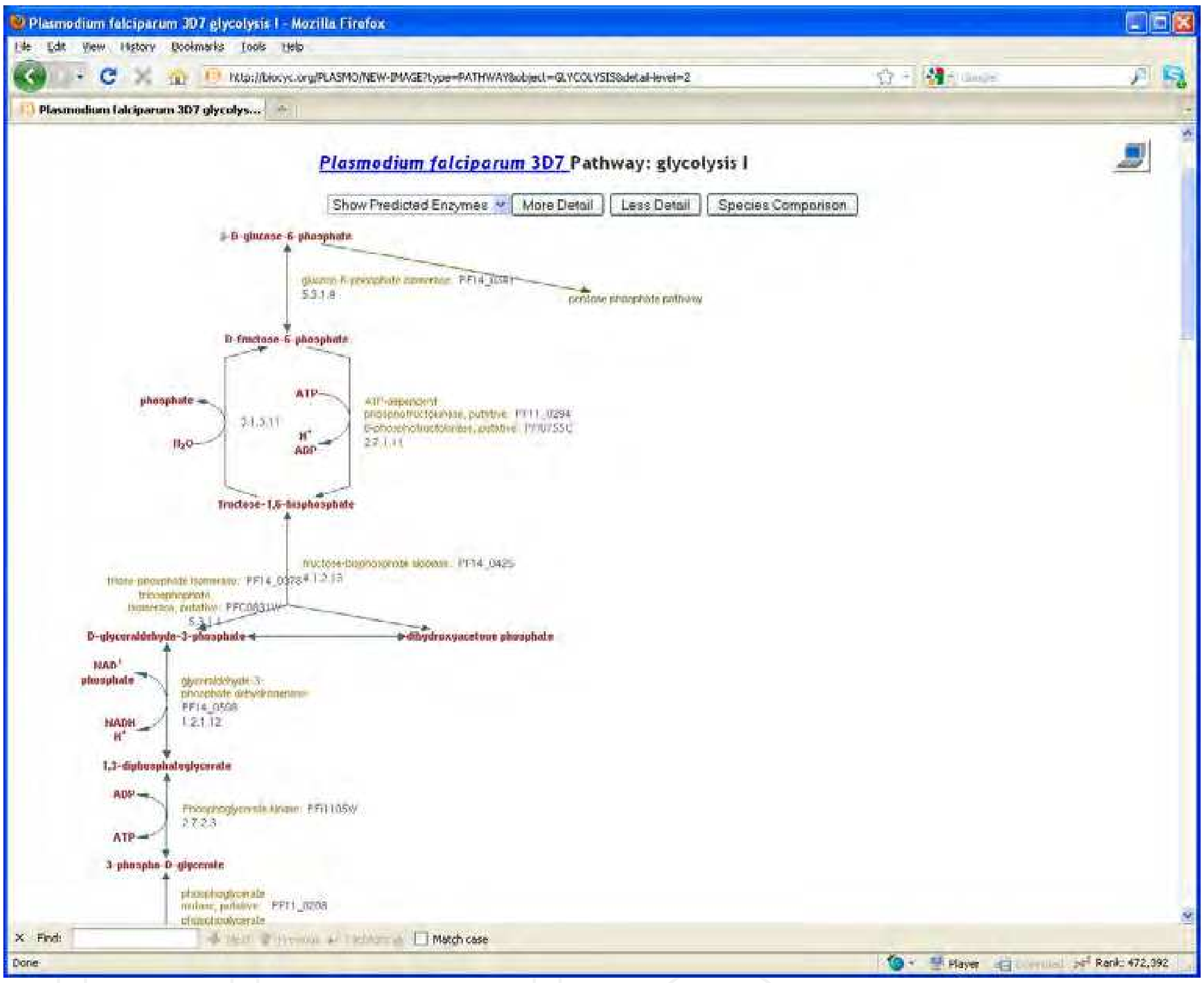
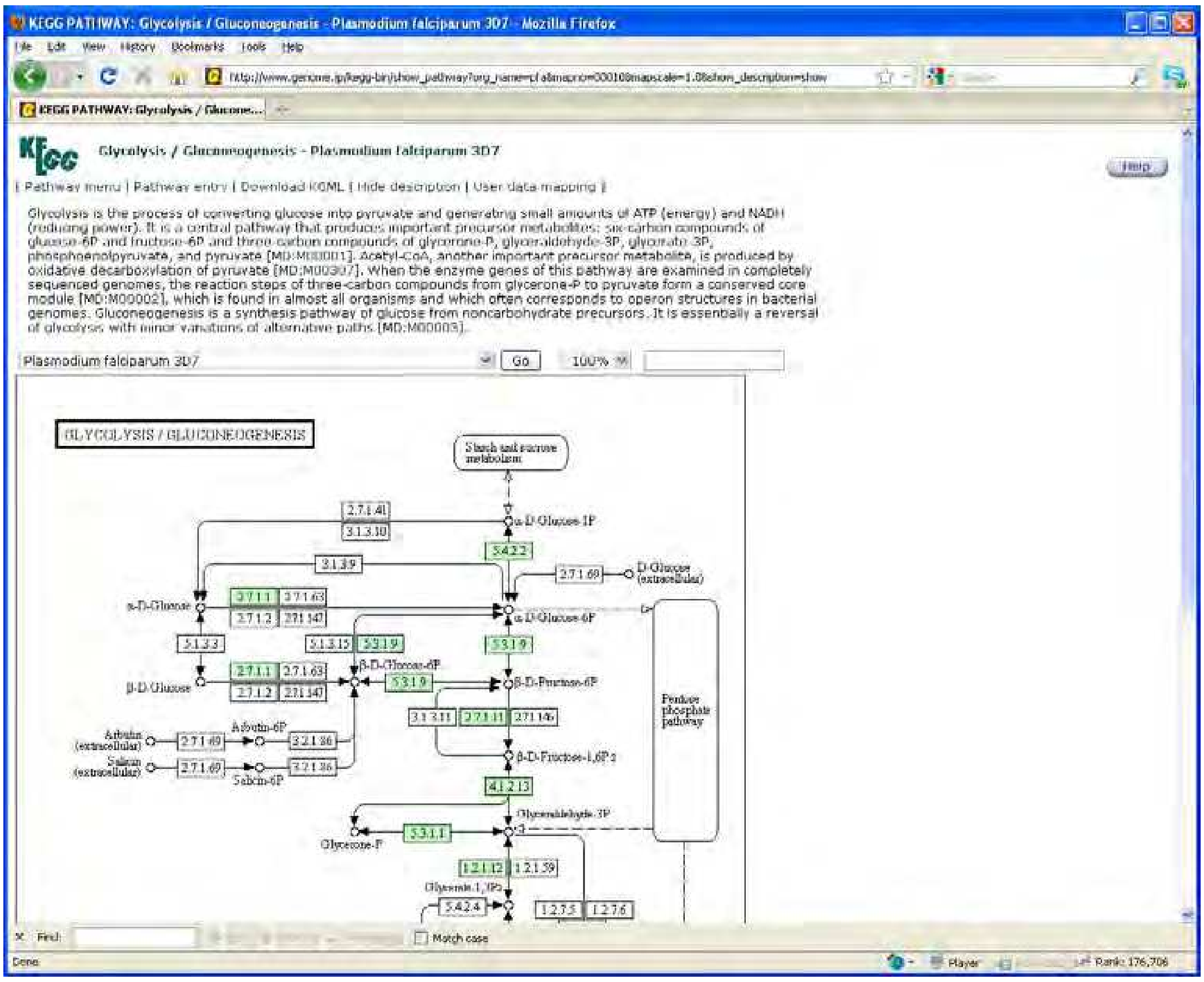

Plasmodium falciparum 3D7 glycolysis I - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://biocyc.org/PLASMO/NEW-IMAGE?type=PATHWAY&object=GLYCOLYSIS&detail-level=2
Plasmodium falciparum 3D7 glycolys...
Plasmodium falciparum 3D7 Pathway: glycolysis I
Show Predicted Enzymes
More Detail
Less Detail
Species Comparison
β-D-glucose-6-phosphate
glucose-6-phosphate isomerase: PF14_0341
5.3.1.9
pentose phosphate pathway
D-fructose-6-phosphate
phosphate
H_2O
3.1.3.11
ATP
H^+
ADP
ATP-dependent phosphofructokinase, putative: PF11_0294
6-phosphofructokinase, putative: PFI0755C
2.7.1.11
fructose-1,6-bisphosphate
fructose-bisphosphate aldolase: PF14_0425
4.1.2.13
triose-phosphate isomerase: PF14_0378
triosephosphate isomerase, putative: PFC0831W
5.3.1.1
D-glyceraldehyde-3-phosphate
dihydroxyacetone phosphate
NAD^+
phosphate
glyceraldehyde-3-phosphate dehydrogenase
PF14_0598
1.2.1.12
NADH
H^+
1,3-diphosphateglycerate
ADP
phosphoglycerate kinase: PFI1105W
2.7.2.3
ATP
3-phospho-D-glycerate
phosphoglycerate mutase, putative: PF11_0208
phosphoglycerate
Find:
Next
Previous
Highlight all
☐ Match case
Done
Player
Rank: 472,392

Fig. 4. The glycolysis pathway, as shown in PlasmoCyc. Enzymes, chemical compounds and co-factors are shown at this level. Multiple levels of detail may be selected for visualization.

Additionally, KEGG provides a series of analysis functionalities such as tools for mapping molecules to pathways and generating graphical representations, predicting metabolic pathways, annotation tools, sequence similarity searching and chemical similarity searching. A particularly useful feature is the ability to display gene expression data onto the KEGG pathways using the KEGG expression resource. A downloadable application is also provided for detailed analysis of gene expression data together with KEGG pathways and KEGG genomes. Many of the KEGG features may be accessed as a web service using the provided WSDL specifications. The KEGG pathways are also available for download in XML format.



KEGG PATHWAY: Glycolysis / Gluconeogenesis - Plasmodium falciparum 3D7 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.genome.jp/kegg-bin/show_pathway?org_name=pfa&mapno=00010&mapscale=1.0&show_description=show
KEGG PATHWAY: Glycolysis / Glucone...
Help
Find:
Match case
Done
Player
Rank: 176,706

Glycolysis / Gluconeogenesis - Plasmodium falciparum 3D7

| Pathway menu | Pathway entry | Download KGML | Hide description | User data mapping |

Glycolysis is the process of converting glucose into pyruvate and generating small amounts of ATP (energy) and NADH (reducing power). It is a central pathway that produces important precursor metabolites: six-carbon compounds of glucose-6P and fructose-6P and three-carbon compounds of glycerone-P, glyceraldehyde-3P, glycerate-3P, phosphoenolpyruvate, and pyruvate [MD:M00001]. Acetyl-CoA, another important precursor metabolite, is produced by oxidative decarboxylation of pyruvate [MD:M00307]. When the enzyme genes of this pathway are examined in completely sequenced genomes, the reaction steps of three-carbon compounds from glycerone-P to pyruvate form a conserved core module [MD:M00002], which is found in almost all organisms and which often corresponds to operon structures in bacterial genomes. Gluconeogenesis is a synthesis pathway of glucose from noncarbohydrate precursors. It is essentially a reversal of glycolysis with minor variations of alternative paths [MD:M00003].

Plasmodium falciparum 3D7 Go 100%

Fig. 5. The glycolysis pathway, as shown in KEGG. EC numbers and chemical compounds names are shown.

4. Protein-protein interaction databases

4.1 PlasmoMAP

URL: http://www.cbil.upenn.edu/plasmoMAP

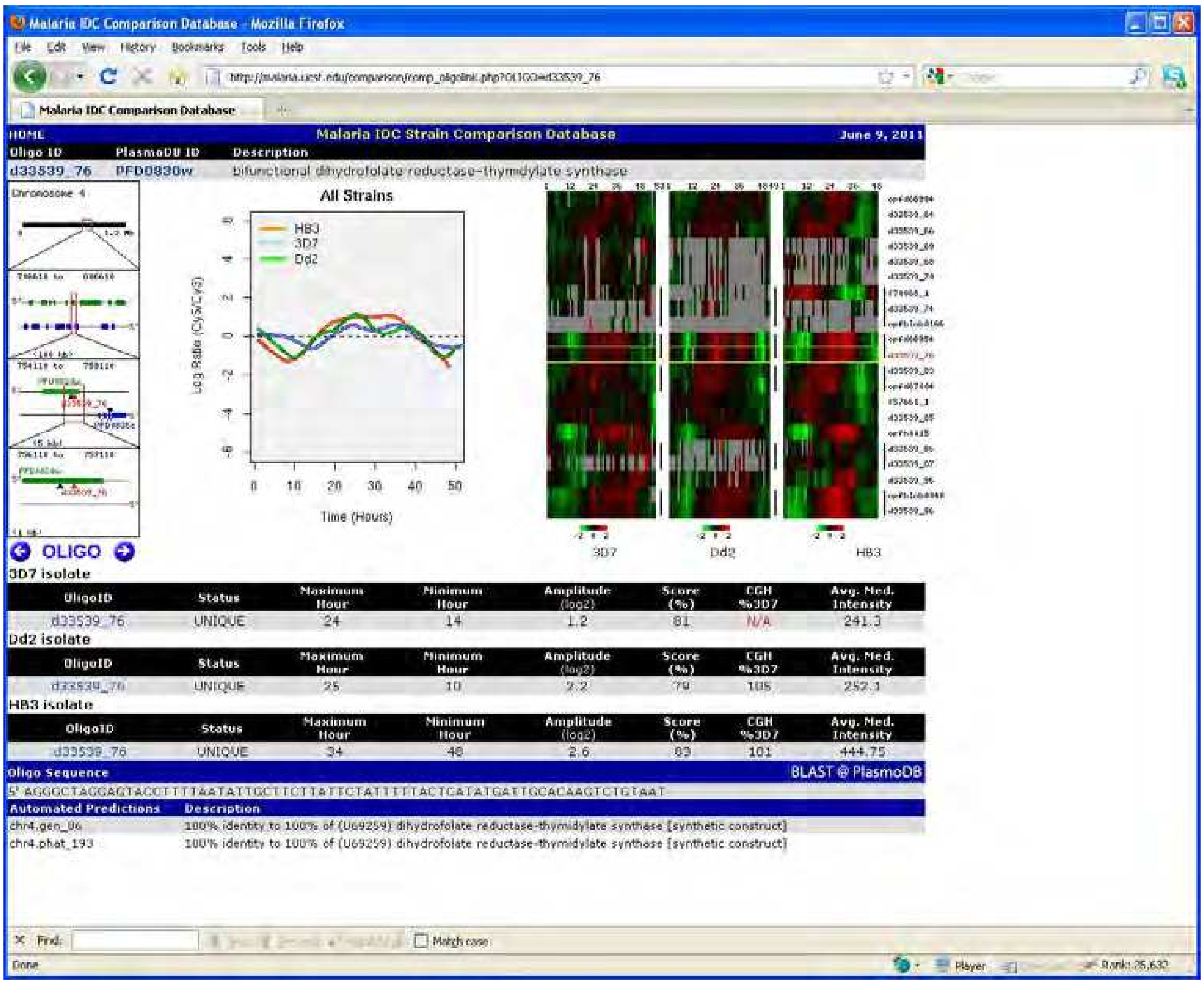
PlasmoMAP is focussed at functional interactions between proteins in *P. falciparum* (Date & Stoeckert, 2006), with the goal of eventually illustrating a complete interactome. It is based on the reconstruction of functional genomics and computational data using a Bayesian framework. The interaction network covers around 68% of the parasite genome and infers information for nearly 2000 uncharacterized proteins. PlasmoMAP contains data for *P. falciparum* strains HB3, 3D7 and Dd2. The resource allows the user to access the data by either generating a complete protein-protein interaction dataset at a specified cut-off value, or by specifying a protein name for which to generate interaction partners. The PlasmoMAP network data is made available for download in LGL format, so that the networks may be viewed using the LGLView application. Additionally, the raw functional genomics data may be downloaded, as well as various subsets of the interaction data by confidence, by KEGG category, by GO category and by linkages detected in other apicomplexan organisms.

5. Gene expression databases

5.1 Malaria IDC strain comparison database

URL: http://malaria.ucsf.edu

The IDC site presents the data from the DeRisi lab studies on the transcriptome analysis of *P. falciparum* (Bozdech *et al.*, 2003; Llinás *et al.*, 2006). Queries may be executed using a variety of terms, including PlasmoDB ID, ORF description, oligonucleotide ID, automated prediction, automated description, GO annotation, GO number, common gene name, functional group, chromosome or by providing a list of ORFs or oligonucleotides. Strain, time, Fourier, amplitude and CGH constraints may be set for queries. In this example of viewing the expression of a specific gene, a query was performed for the PlasmoDB ID PFD0830w (bifunctional dihydrofolate reductase-thymidylate) and the expression data for the d33539_76 oligonucleotide was visualized for the 3D7 isolate (Fig 6). The chromosomal



Malaria IDC Comparison Database - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://malaria.ucsf.edu/comparison/comp_oligolink.php?OLIGO=d33539_76
Malaria IDC Comparison Database
HOME
Malaria IDC Strain Comparison Database
June 9, 2011
Chromosome 4
All Strains
HB3
3D7
Dd2
Log Ratio (Cy5/Cy3)
Time (Hours)
0
10
20
30
40
50
3D7
Dd2
HB3
OLIGO
Find:
Match case
Done

Oligo ID	PlasmoDB ID	Description
d33539_76	PFD0830w	bifunctional dihydrofolate reductase-thymidylate synthase

3D7 isolate

OligoID	Status	Maximum Hour	Minimum Hour	Amplitude (log2)	Score (%)	CGH %3D7	Avg. Med. Intensity
d33539_76	UNIQUE	24	14	1.2	81	N/A	241.3

Dd2 isolate

OligoID	Status	Maximum Hour	Minimum Hour	Amplitude (log2)	Score (%)	CGH %3D7	Avg. Med. Intensity
d33539_76	UNIQUE	25	10	2.2	79	105	252.1

HB3 isolate

OligoID	Status	Maximum Hour	Minimum Hour	Amplitude (log2)	Score (%)	CGH %3D7	Avg. Med. Intensity
d33539_76	UNIQUE	34	48	2.6	83	101	444.75

Oligo Sequence BLAST @ PlasmoDB

5' AGGGCTAGGAGTACCTTTTAATATTGCTTCTTATTCTATTTTTACTCATATGATTGCACAAGTCTGTAAT

Automated Predictions	Description
chr4.gen_06	100% identity to 100% of (U69259) dihydrofolate reductase-thymidylate synthase [synthetic construct]
chr4.phat_193	100% identity to 100% of (U69259) dihydrofolate reductase-thymidylate synthase [synthetic construct]

Fig. 6. The expression profile of PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase from *P. falciparum*) visualized in the Malaria IDC Strain Comparison Database as measured using the d33539_79 oligonucleotide probe. The chromosomal position is shown on the left. The expression of the gene over time is visible as a graph for the different parasite trains, and a heat map of expression measured by other probes are shown on the right. Detailed information is provided at the bottom.

location, log-ratio of expression over time and strain comparison heat map are shown together with a table providing more detailed time, median and scoring information. When a search is performed using specific constraints over the complete dataset, matching genes are shown in a tabular fashion, together with a miniaturized heatmap of the expression profile. The IDC data is also made available for download in normalized as well as non-normalized formats. Furthermore, the raw image files from the study may also be obtained.

6. Functional prediction databases

6.1 PlasmoDraft

URL: http://www.atgc-montpellier.fr/PlasmoDraft

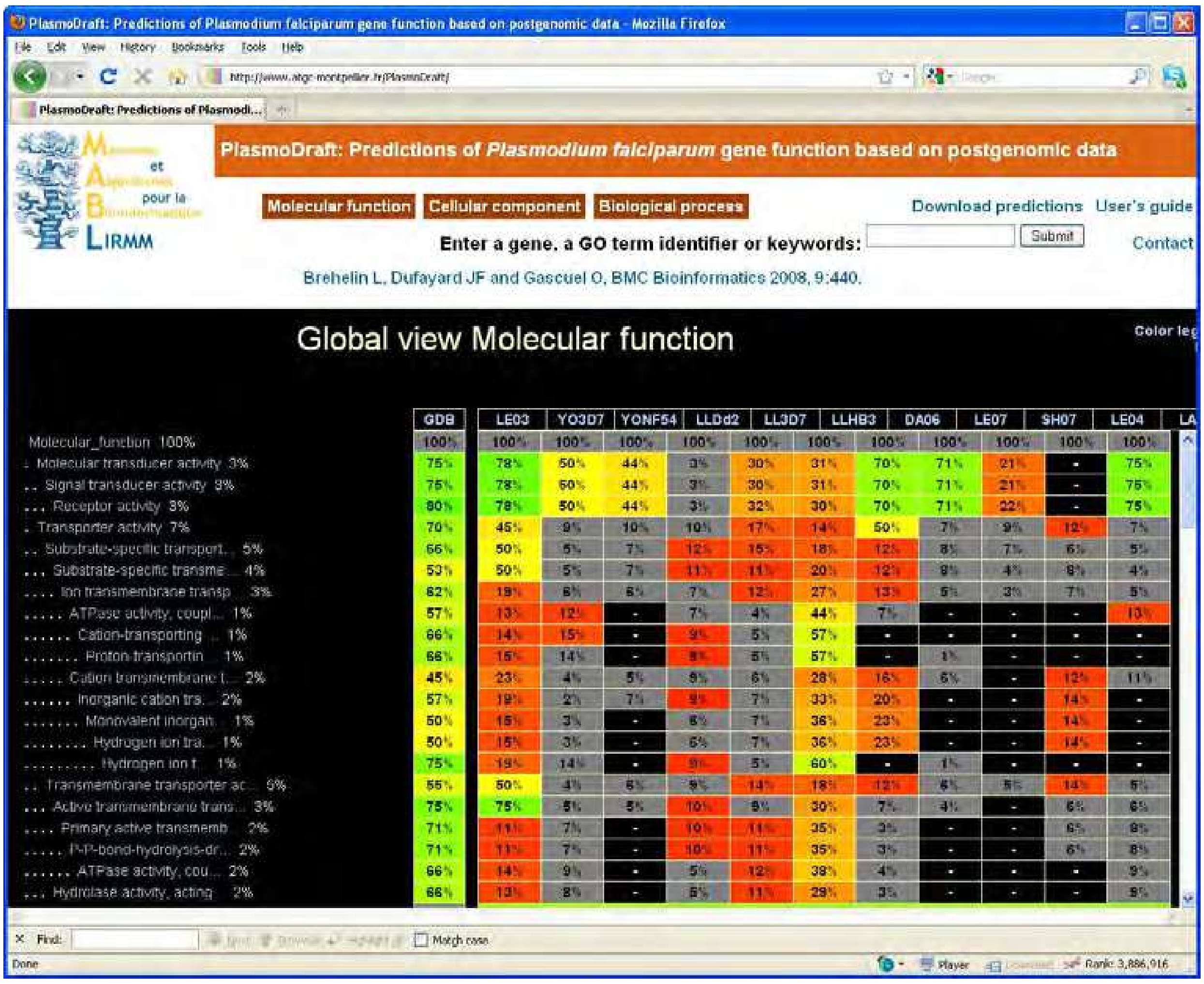
The PlasmoDraft database predicts GO terms for genes from *P. falciparum*. It is based on a Guilt By Association (GBA) predictor named Gonna, measuring the profile of a gene's similarity using data from transcriptome, proteome and interactome studies to genes in the GeneDB database (Bréhélin *et al.*, 2008). The database may be viewed globally for each gene ontology category (Molecular Function, Biological Process and Cellular Component). The GO term is displayed, followed by its parameters for a series of parasite strains. These include prior probability (indicating term frequency), the best global degree of belief (GDB) and its confidence (TDR: True Discovery Rate). These are indicated using a color code varying from green to red (Figure 7). The database may also be searched using a gene, a GO term identifier or a keyword. This approach to functional prediction is complementary to the classical sequence based-approaches. Around 60% of genes lacking annotations are included, in addition to genes already annotated in GeneDB. Particularly useful is the ability to identify GO terms attached to a specific gene, and also genes attached to a specific GO term.

7. Protein structure databases

7.1 PDB

URL: http://www.pdb.org

Although not malaria-specific, the PDB database is the global repository for all protein structure information, derived primarily from X-ray crystallography and NMR studies (Bernstein *et al.*, 1977). At the time of writing, there were 59 structures from Plasmodia, of which some were duplicates of the same protein with different ligands or with single / multiple mutations. The PDB database may be searched using an extensive range of criteria, but the main approaches are using a text term search or performing a BLAST search using a protein sequence to find homologous proteins. The results view of a protein structure includes a preview of the protein structure, literature information, a molecular description, information about related PDB entries and information about associated ligands (Figure 8). The structure may be interactively viewed using JMol (http://www.jmol.org). The structural and sequence data may be downloaded in a range of formats for further investigation.



PlasmoDraft: Predictions of Plasmodium falciparum gene function based on postgenomic data - Mozilla Firefox

PlasmoDraft: Predictions of *Plasmodium falciparum* gene function based on postgenomic data

Enter a gene, a GO term identifier or keywords:

Brehelin L, Dufayard JF and Gascuel O, BMC Bioinformatics 2008, 9:440.

Global view Molecular function

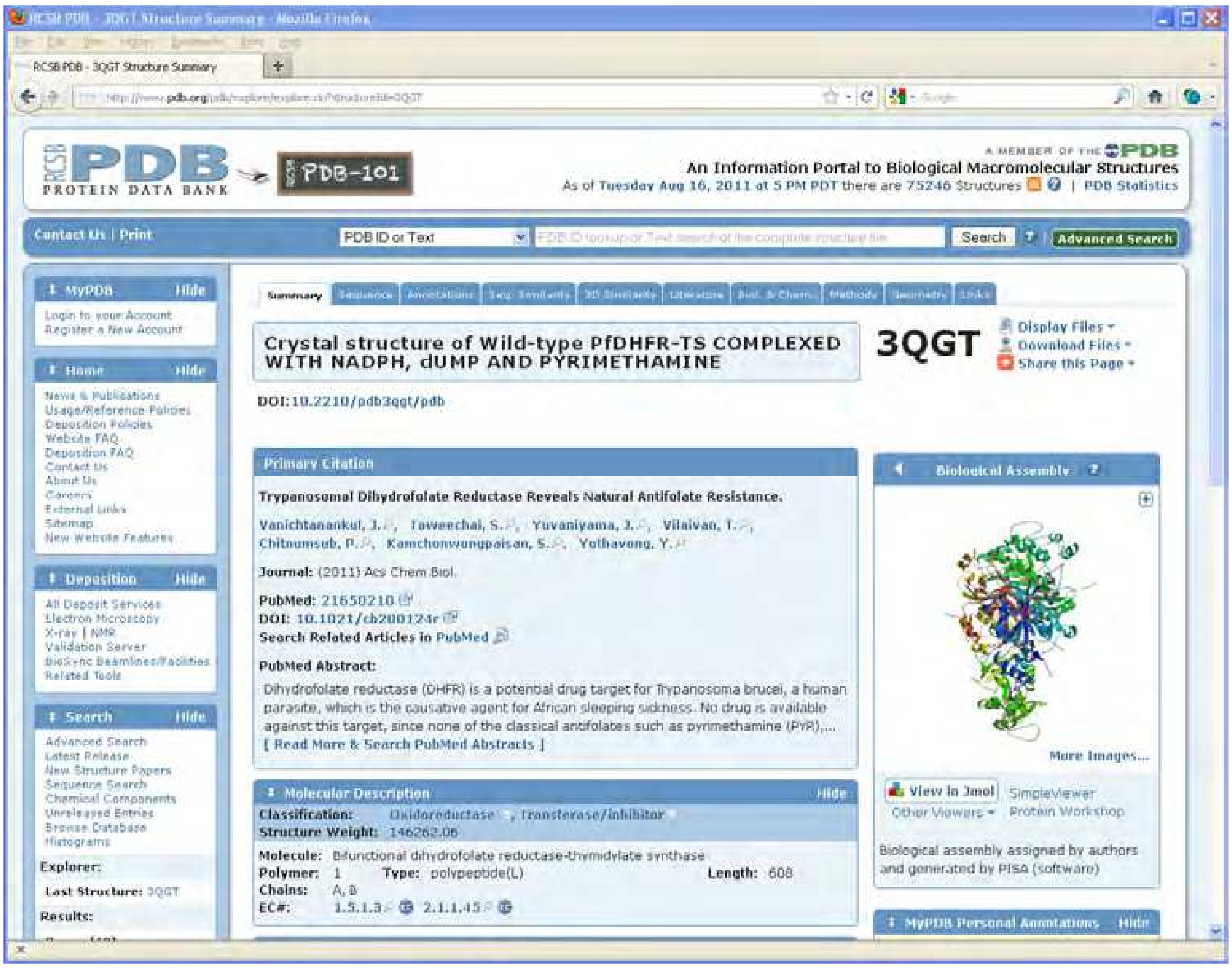
	GDB	LE03	YO3D7	YONF54	LLDd2	LL3D7	LLHB3	DA06	LE07	SH07	LE04	LA
Molecular_function 100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
. Molecular transducer activity 3%	75%	78%	50%	44%	3%	30%	31%	70%	71%	21%	-	75%
.. Signal transducer activity 3%	75%	78%	50%	44%	3%	30%	31%	70%	71%	21%	-	75%
... Receptor activity 3%	80%	78%	50%	44%	3%	32%	30%	70%	71%	22%	-	75%
. Transporter activity 7%	70%	45%	9%	10%	10%	17%	14%	50%	7%	9%	12%	7%
.. Substrate-specific transport... 5%	66%	50%	5%	7%	12%	15%	18%	12%	8%	7%	6%	5%
... Substrate-specific transme... 4%	53%	50%	5%	7%	11%	11%	20%	12%	9%	4%	8%	4%
.... Ion transmembrane transp... 3%	62%	19%	6%	6%	7%	12%	27%	13%	5%	3%	7%	5%
..... ATPase activity, coupl... 1%	57%	13%	12%	-	7%	4%	44%	7%	-	-	-	13%
...... Cation-transporting ... 1%	66%	14%	15%	-	9%	5%	57%	-	-	-	-	-
.... Proton-transportin... 1%	66%	15%	14%	-	9%	5%	57%	-	1%	-	-	-
..... Cation transmembrane t... 2%	45%	23%	4%	5%	9%	6%	28%	16%	6%	-	12%	11%
...... Inorganic cation tra... 2%	57%	19%	2%	7%	9%	7%	33%	20%	-	-	14%	-
.... Monovalent inorgan... 1%	50%	15%	3%	-	6%	7%	36%	23%	-	-	14%	-
.... Hydrogen ion tra... 1%	50%	15%	3%	-	6%	7%	36%	23%	-	-	14%	-
.... Hydrogen ion t... 1%	75%	19%	14%	-	9%	5%	60%	-	1%	-	-	-
.. Transmembrane transporter ac... 5%	55%	50%	4%	6%	9%	14%	18%	12%	6%	5%	14%	5%
... Active transmembrane trans... 3%	75%	75%	5%	5%	10%	9%	30%	7%	4%	-	6%	6%
.... Primary active transmemb... 2%	71%	11%	7%	-	10%	11%	35%	3%	-	-	6%	8%
..... P-P-bond-hydrolysis-dr... 2%	71%	11%	7%	-	10%	11%	35%	3%	-	-	6%	8%
...... ATPase activity, cou... 2%	66%	14%	9%	-	5%	12%	38%	4%	-	-	-	9%
... Hydrolase activity, acting... 2%	66%	13%	8%	-	6%	11%	29%	3%	-	-	-	8%

Fig. 7. A global view for a series of parasite strains based on the GO Molecular Function classification in PlasmoDraft. GO categories are listed on the left of the table, and the results are shown in columns for a series of different parasite strains.

7.2 TDI Kernel

URL: http://tropicaldisease.org/kernel

The TDI Kernel strongly supports the Open Source approach to drug discovery, and is aimed at the prediction of binding sites and ligand interactions through homology modeling approaches (Ortí *et al.*, 2009). Targets may be searched by organism, which includes *C. hominis*, *C. parvum*, *L. major*, *M. leprae*, *M. tuberculosis*, *P. falciparum*, *P. vivax*, *T. brucei*, *T. cruzi* and *T. gondii*. Further search criteria include keywords, UniProt IDs, DrugBank IDs, PDB Ligand ID and homology modeling parameters. At the time of writing, 28 templates had been analyzed for *P. falciparum*. The detailed view of a target shows the templates used for homology modeling, provides the generated models and shows the ligand predictions for the selected target (Figure 9). Each of these may be investigated further through links to PDB, MSD (http://www.ebi.ac.uk/msd) and DrugBank (Wishart *et al.*, 2008). For each prediction, the result may be expanded to further show a figure of the ligand bound in the active site, drug category information for the ligand, and current uses of the drug.



http://www.pdb.org/pdb/explore/explore.do?structureId=3QGT

An Information Portal to Biological Macromolecular Structures
As of Tuesday Aug 16, 2011 at 5 PM PDT there are 75246 Structures | PDB Statistics

Crystal structure of Wild-type PfDHFR-TS COMPLEXED WITH NADPH, dUMP AND PYRIMETHAMINE

3QGT

DOI:10.2210/pdb3qgt/pdb

Trypanosomal Dihydrofolate Reductase Reveals Natural Antifolate Resistance.

Vanichtanankul, J., Taweechai, S., Yuvaniyama, J., Vilaivan, T., Chitnumsub, P., Kamchonwongpaisan, S., Yuthavong, Y.

Journal: (2011) Acs Chem.Biol.

PubMed: 21650210
DOI: 10.1021/cb200124r
Search Related Articles in PubMed

PubMed Abstract:
Dihydrofolate reductase (DHFR) is a potential drug target for Trypanosoma brucei, a human parasite, which is the causative agent for African sleeping sickness. No drug is available against this target, since none of the classical antifolates such as pyrimethamine (PYR),...
[Read More & Search PubMed Abstracts]

Classification: Oxidoreductase, Transferase/inhibitor
Structure Weight: 146262.06

Molecule: Bifunctional dihydrofolate reductase-thymidylate synthase
Polymer: 1 Type: polypeptide(L) Length: 608
Chains: A, B
EC#: 1.5.1.3 2.1.1.45

Biological assembly assigned by authors and generated by PISA (software)

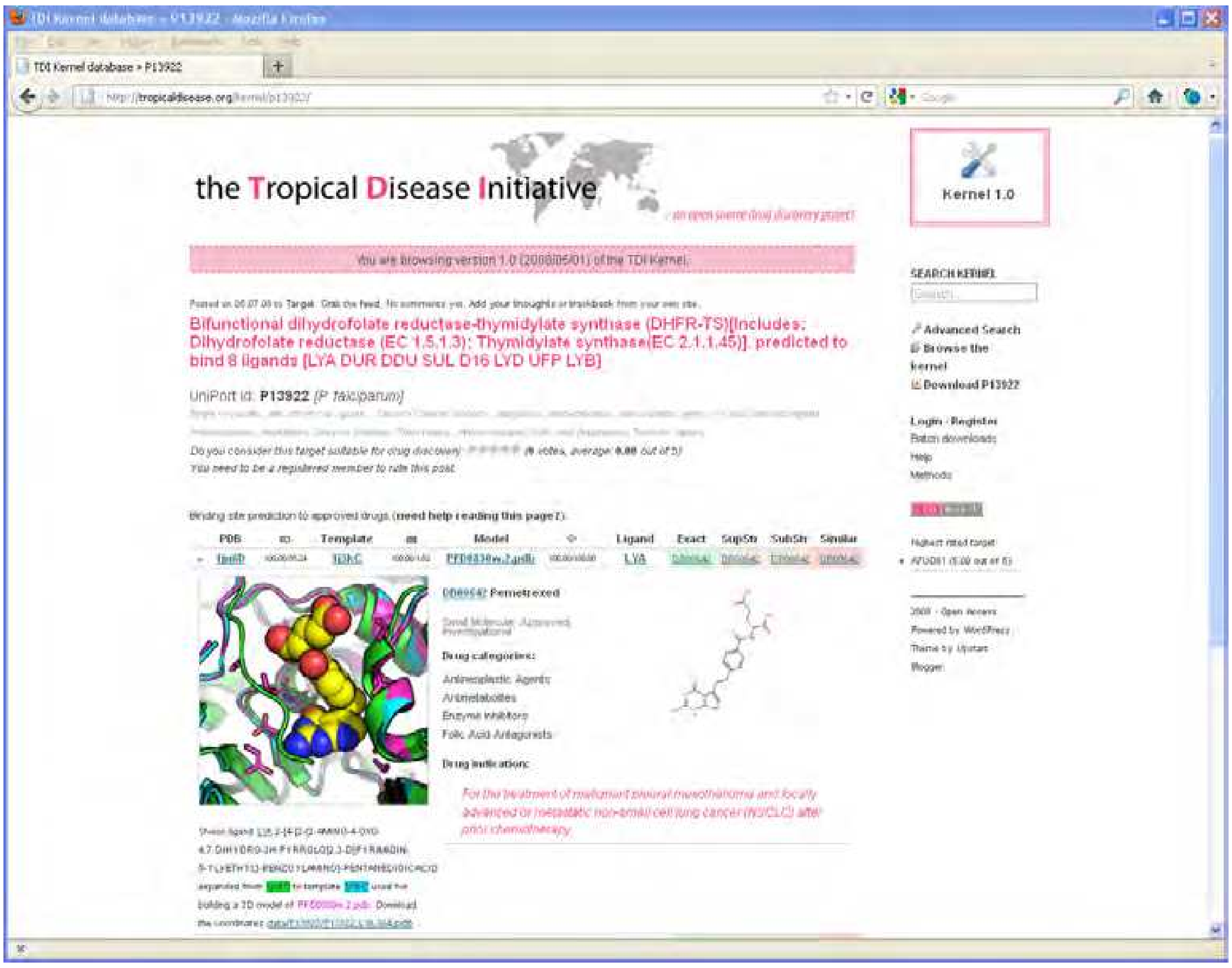
Last Structure: 3QGT

Fig. 8. The PDB results page for a structure of PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase from *P. falciparum*). A summary of the protein's information is shown in the centre, with a structural overview on the right.

7.3 ModBase

URL: http://modbase.compbio.ucsf.edu

ModBase is an extensive collection of homology models for a variety of species, including *P. falciparum* and *P. vivax* (Pieper *et al.*, 2009). The homology models have been constructed using ModPipe, a pipeline which uses mainly PSI-BLAST (Altschul *et al.*, 1997) and Modeler (Eswar *et al.*, 2006) for model building. At the time of writing, ModBase contained models for 2599 *P. falciparum* and 2359 *P. vivax* proteins. Searches may be performed using text terms or protein sequence, and results may be displayed as either a model overview, detail regarding the model or a sequence overview. The detailed view contains information about the model coverage of the protein sequence, model quality parameters, graphical representations of the protein structure and a wide range of external links to other relevant databases (Figure 10). Where available, SNP information and ligand binding sites may also be viewed.



TDI Kernel database » P13922

http://tropicaldisease.org/kernel/p13922/

the Tropical Disease Initiative

Kernel 1.0

You are browsing version 1.0 (2008/05/01) of the TDI Kernel.

Bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS)[Includes: Dihydrofolate reductase (EC 1.5.1.3); Thymidylate synthase(EC 2.1.1.45)]. predicted to bind 8 ligands [LYA DUR DDU SUL D16 LVD UFP LYB]

UniPort id: **P13922** *(P. falciparum)*

Do you consider this target suitable for drug discovery? (0 votes, average: 0.00 out of 5)

You need to be a registered member to rate this post.

Binding site prediction to approved drugs (**need help reading this page?**)

PDB	ID	Template		Model		Ligand	Exact	SupStr	SubStr	Similar
1juiD		1J3KC		PFD0830w.2.pdb		LYA				

DB00642 Pemetrexed

Drug categories:

Antineoplastic Agents

Antimetabolites

Enzyme Inhibitors

Folic Acid Antagonists

Drug indication:

For the treatment of malignant pleural mesothelioma and locally advanced or metastatic non-small cell lung cancer (NSCLC) after prior chemotherapy.

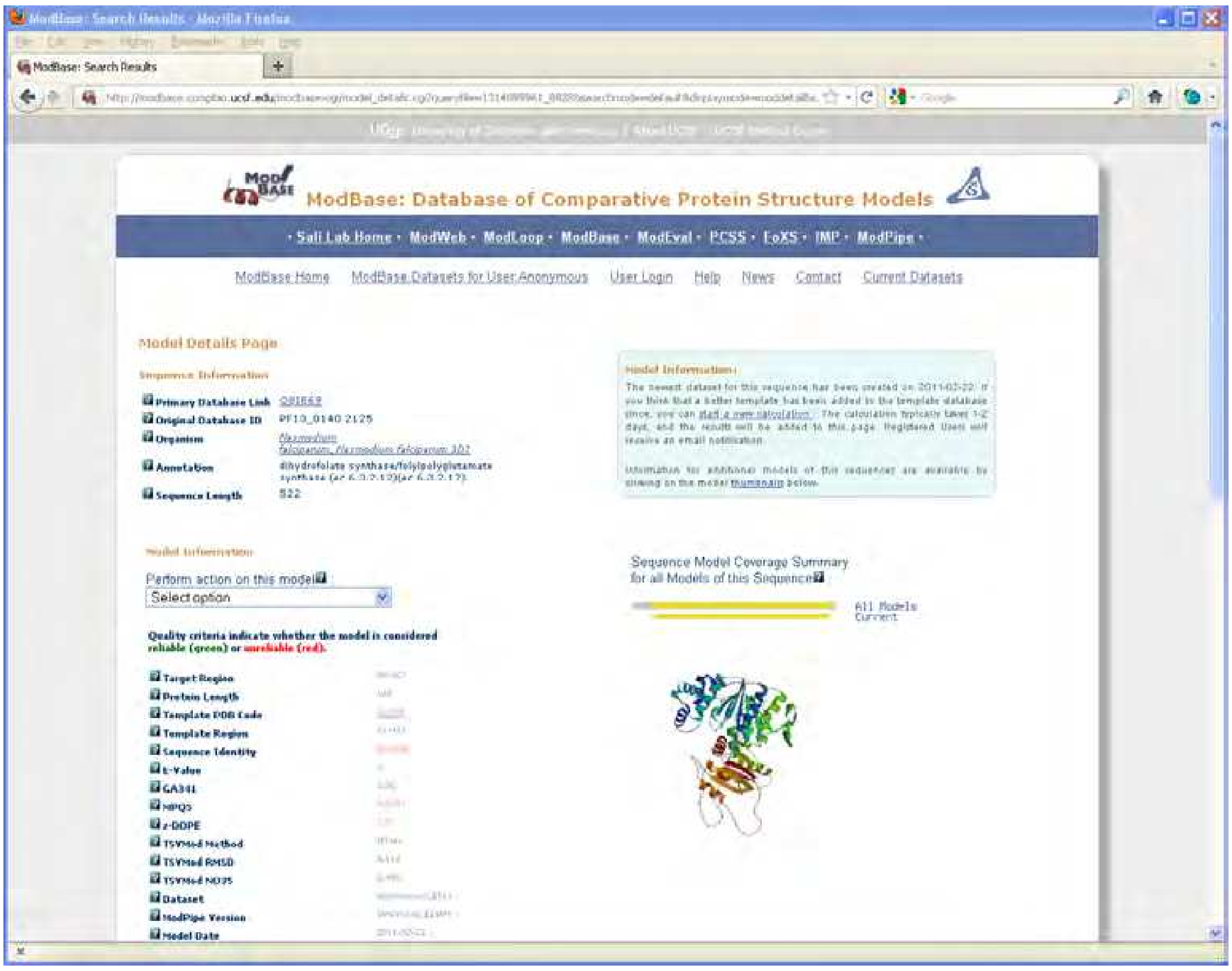
Fig. 9. The result view for PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase from *P. falciparum*) with LYA in the TDI Kernel resource. A structural view of the ligand in binding site is shown, as well as additional information about the protein and ligand.

8. Literature databases

8.1 The malaria literature database

URL: http://carrier.gnf.org/publications/Py

The Malaria Literature Database is compiled by the Genomics Institute of the Novartis Research Foundation. Whereas most literature searches focus only on abstracts, Google Scholar and Scirus is used to search in the body of Open Access papers for the presence of gene names. Gene name - publication pairs are then collected and indexed. Additionally, searches may be performed for genes, orthologs and published papers. Gene results are displayed with PlasmoDB links, and ortholog results are shown. Literature results are shown together with PubMed PMID links for easy access to the abstracts and papers. It is also possible to download all data in batch format.



ModBase: Search Results

ModBase: Database of Comparative Protein Structure Models

· Sali Lab Home · ModWeb · ModLoop · ModBase · ModEval · PCSS · FoXS · IMP · ModPipe ·
ModBase Home
ModBase Datasets for User Anonymous
User Login
Help
News
Contact
Current Datasets

Model Details Page

Sequence Information

Primary Database Link	Q8I5S9
Original Database ID	PF10_0140 2125
Organism	Plasmodium falciparum, Plasmodium falciparum 3D7
Annotation	dihydrofolate synthase/folylpolyglutamate synthase (ec 6.3.2.12)(ec 6.3.2.17)
Sequence Length	522

Model Information:

The newest dataset for this sequence has been created on 2011-02-22. If you think that a better template has been added to the template database since, you can start a new calculation. The calculation typically takes 1-2 days, and the results will be added to this page. Registered Users will receive an email notification.

Information for additional models of this sequence are available by clicking on the model thumbnail below.

Model Information

Perform action on this model
Select option

Sequence Model Coverage Summary for all Models of this Sequence

Quality criteria indicate whether the model is considered reliable (green) or unreliable (red).

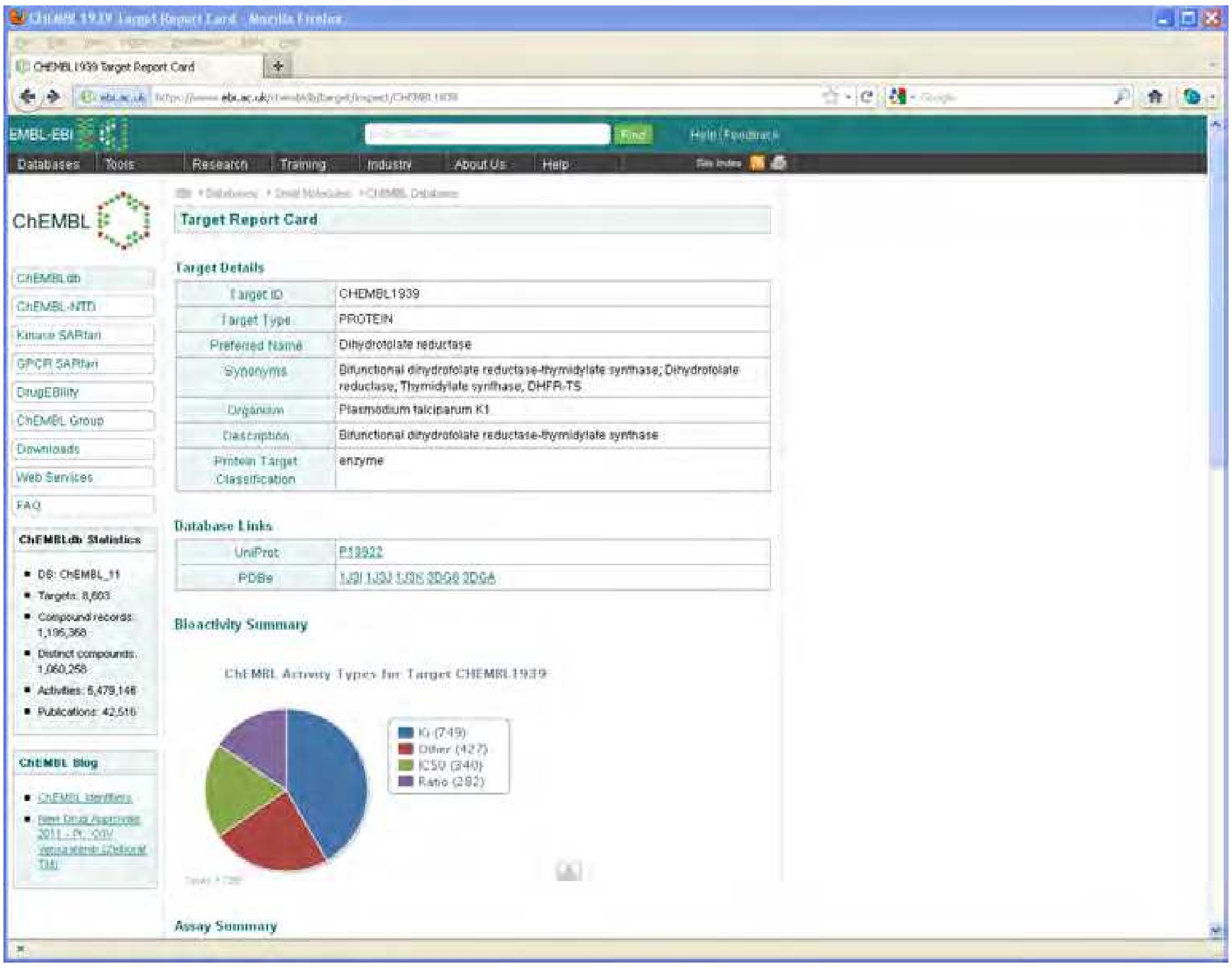
Target Region	[illegible]
Protein Length	[illegible]
Template PDB Code	[illegible]
Template Region	[illegible]
Sequence Identity	[illegible]
E-Value	[illegible]
GA341	[illegible]
MPQS	[illegible]
z-DOPE	[illegible]
TSVMod Method	[illegible]
TSVMod RMSD	[illegible]
TSVMod NO35	[illegible]
Dataset	[illegible]
ModPipe Version	[illegible]
Model Date	[illegible]

Fig. 10. A detailed view of a homology model of PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase from *P. falciparum*). Model parameters are shown on the left, with an alignment overview and graphical representation of the protein structure on the right.

9. Drug discovery databases

9.1 ChEMBL

The ChEMBL database is primarily a database of bioactive small molecules, and provides access to a wide range of targets and chemical compounds, including industry-generated datasets such as the GSK TCAMS malaria dataset and the Novartis-GNF malaria box set in the ChEMBL-NTD section (Overington, 2009). The targets section may be searched using either a keyword, protein sequence or a ChEMBL ID. The results for a specific target include bioactivity data, assay categories and parameters available for the target and a chemical classification of the related compounds by molecular weight, logP and PSA (Figure 11).



Target Report Card

Target Details

Target ID	CHEMBL1939
Target Type	PROTEIN
Preferred Name	Dihydrofolate reductase
Synonyms	Bifunctional dihydrofolate reductase-thymidylate synthase; Dihydrofolate reductase; Thymidylate synthase; DHFR-TS
Organism	Plasmodium falciparum K1
Description	Bifunctional dihydrofolate reductase-thymidylate synthase
Protein Target Classification	enzyme

Database Links

UniProt	P13922
PDBe	1J3I 1J3J 1J3K 3DG8 3DGA

ChEMBLdb Statistics

- DB: ChEMBL_11
- Targets: 8,603
- Compound records: 1,196,368
- Distinct compounds: 1,060,258
- Activities: 5,479,146
- Publications: 42,516

Bioactivity Summary

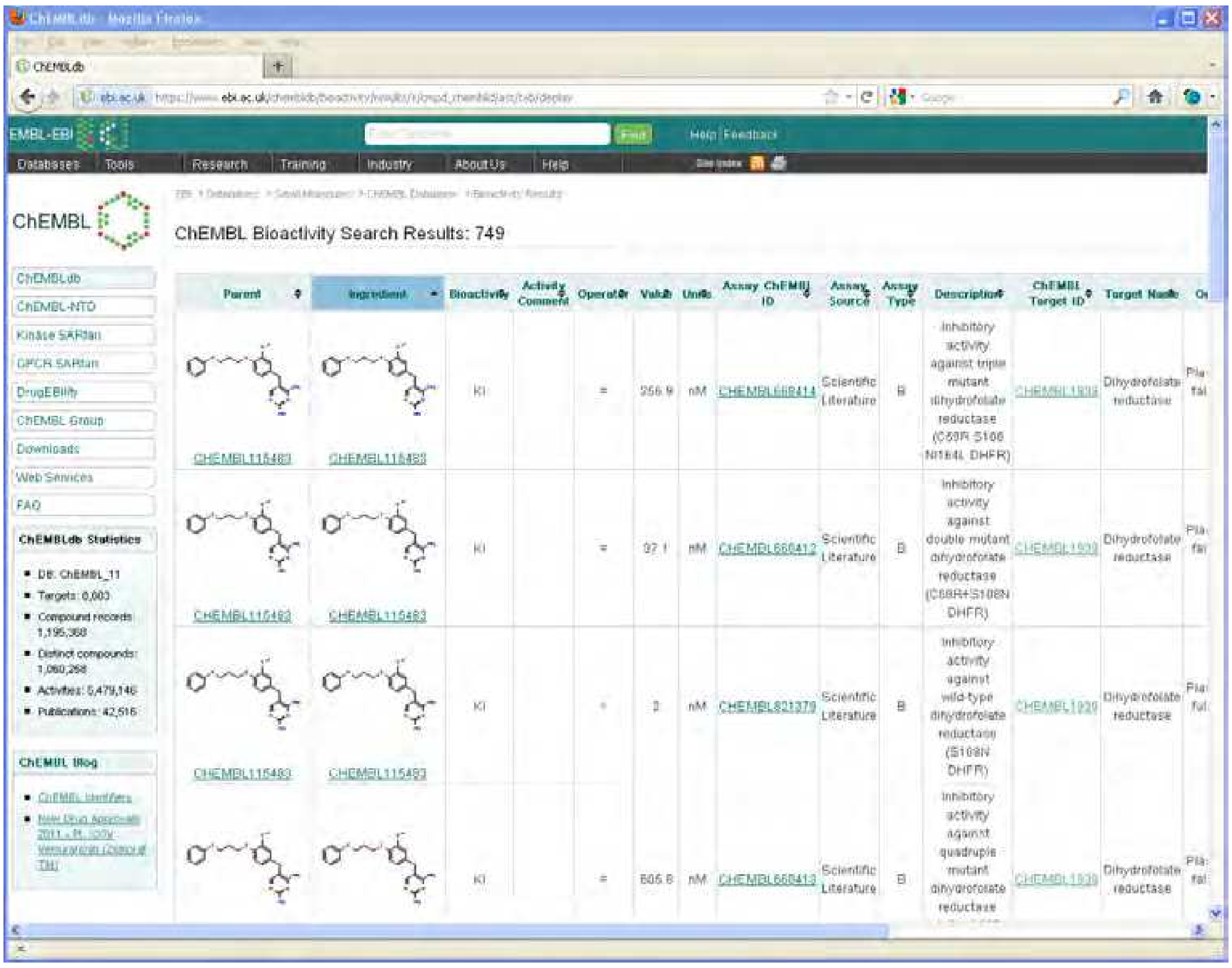
ChEMBL Activity Types for Target CHEMBL1939

Assay Summary

Fig. 11. The result view for PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase from *P. falciparum*) in ChEMBL. The protein's details are shown at the top, with database links followed by a graphical summary of available binding data at the bottom.

These are presented as a graphical summary view, where each section may be explored further. The next level of results is a table showing the compound structure, bioactivity, assay type, assay source, results and literature reference (Figure 12). Links between proteins and compounds are established through manual literature curation.

Chemicals may be searched using text-based queries or chemical structures. Results are returned in a tabular format with chemical structures and detailed chemical properties. When a compound is selected for further exploration, clinical trials relevant to the compound, bioactivity, assay and protein target information are also shown together with links to external databases. A search functionality for assays is also available. The ChEMBL database additionally contains specific resource areas for kinases (Kinase SARfari) and for G protein-coupled receptors (GPCR SARfari). It also contains the DrugEBIlity resource focussed on structure-based drugability.



ChEMBL Bioactivity Search Results: 749

ChEMBLdb Statistics

- DB: ChEMBL_11
- Targets: 8,603
- Compound records: 1,195,368
- Distinct compounds: 1,060,258
- Activities: 5,479,146
- Publications: 42,516

ChEMBL Blog

- ChEMBL Identifiers
- New Drug Approvals 2011 - Pt. XXIV Vemurafenib (Zelboraf TM)

Parent	Ingredient	Bioactivity	Activity Comment	Operator	Value	Units	Assay ChEMBL ID	Assay Source	Assay Type	Description	ChEMBL Target ID	Target Name	Or
CHEMBL115483	CHEMBL115483	Ki		=	256.9	nM	CHEMBL668414	Scientific Literature	B	Inhibitory activity against triple mutant dihydrofolate reductase (C59R S108 N164L DHFR)	CHEMBL1939	Dihydrofolate reductase	Pla fal
CHEMBL115483	CHEMBL115483	Ki		=	97.1	nM	CHEMBL668412	Scientific Literature	B	Inhibitory activity against double mutant dihydrofolate reductase (C59R+S108N DHFR)	CHEMBL1939	Dihydrofolate reductase	Pla fal
CHEMBL115493	CHEMBL115493	Ki		=	3	nM	CHEMBL821379	Scientific Literature	B	Inhibitory activity against wild-type dihydrofolate reductase (S108N DHFR)	CHEMBL1939	Dihydrofolate reductase	Pla ful
		Ki		=	605.8	nM	CHEMBL668413	Scientific Literature	B	Inhibitory activity against quadruple mutant dihydrofolate reductase	CHEMBL1939	Dihydrofolate reductase	Pla fal

Fig. 12. A more detailed view for PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase from *P. falciparum*)-related compounds in ChEMBL. The chemical structures of the parent compound (if applicable) and the specific chemical ingredient are shown on the left, followed by columns containing detailed assay and target information.

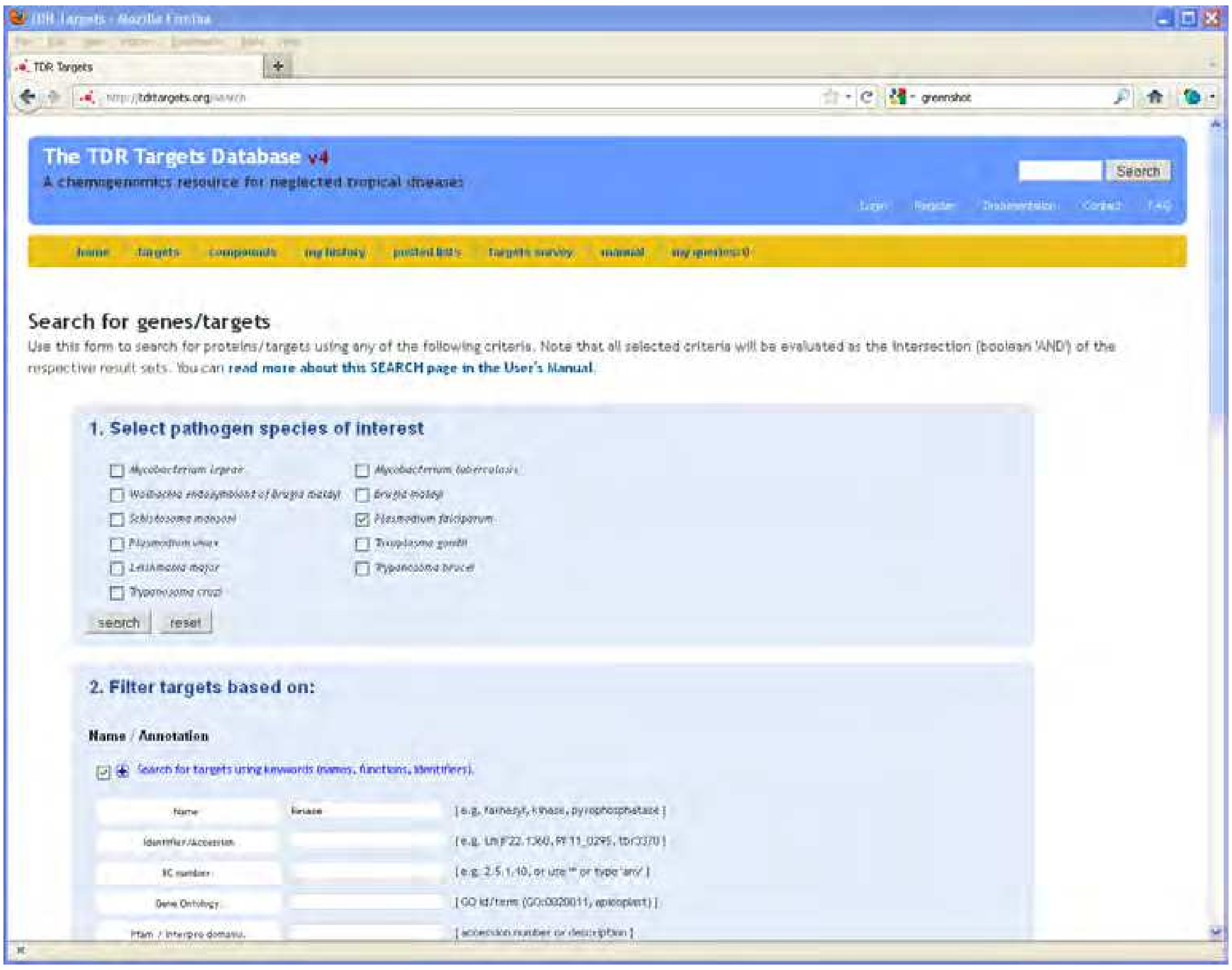
9.2 TDR Targets

URL: http://tdrtargets.org

The TDR Targets database is a comprehensive integrated resource for the selection of drug targets and lead compounds in several different infectious diseases. It is the product of a collaboration of many research groups, and contains very extensive data for *P. falciparum* and *P. vivax*, but also for *T. gondii*, *M. tuberculosis*, *M. leprae*, *T. brucei*, *T. cruzi*, *L. major*, *B. malayi*, *S. mansoni* and *W. bancrofti* (Crowther *et al.*, 2010). It also contains chemical compounds. The TDR Targets database provides pre-compiled lists of targets, but also allows users to perform their own target selection based on a series of molecular properties. Gene information includes a series of fields such as gene ID, gene name, gene product name, exon count, length of gene, length of the protein, molecular weight of protein, isoelectric point of protein, hydrophobicity of proteins, number of transmembrane domains and presence of a signal peptide. Genes are additionally classified into enzymes, transporters and receptors. Functional annotation data includes InterPro domains (Hunter *et al.*, 2009), Pfam domains (Bateman *et al.*, 2004), GO data (Ashburner *et al.*, 2000) and EC numbers.

Structural data contains links to experimentally-derived 3D structures but also to homology models generated as part of the ModBase project. Expression data contains links to a series of microarray experiments. Antigenicity data was generated using the Kolaskar and Tongaonkar method (Kolaskar & Tongaonkar, 1990) as implemented in the EMBOSS antigenic module (Rice *et al.*, 2000). Phyletic distribution information was obtained from the OrthoMCL database (Li *et al.*, 2003). Essentiality data is based on gene knockout and knockdown studies in selected organisms. Drugability data is based on orthology to proteins in the Inpharmatica SAR database which is maintained by the ChEMBL group at the EBI. Drug-to-gene association data is mined from DrugBank and ChEMBL using orthology approaches. Assayability data is primarily obtained from the BRENDA database (Scheer *et al.*, 2011). Literature data originates from PubMed or from curators.

The chemical datasets in TDR Targets is a combination of small molecule data from a variety of sources. This includes properties such as names, synonyms, structure, InChi keys, molecular weight, LogP, hydrogen-bonding donors and acceptors, pharmacological activity and rating according to the Lipinski Rule-of-Five (Lipinski *et al.*, 2001). Bioactivity data is also available.



The TDR Targets Database v4

A chemogenomics resource for neglected tropical diseases

Search for genes/targets

Use this form to search for proteins/targets using any of the following criteria. Note that all selected criteria will be evaluated as the intersection (boolean 'AND') of the respective result sets. You can **read more about this SEARCH page in the User's Manual.**

1. Select pathogen species of interest

□ *Mycobacterium leprae*
□ *Wolbachia endosymbiont of Brugia malayi*
□ *Schistosoma mansoni*
□ *Plasmodium vivax*
□ *Leishmania major*
□ *Trypanosoma cruzi*
□ *Mycobacterium tuberculosis*
□ *Brugia malayi*
☑ *Plasmodium falciparum*
□ *Toxoplasma gondii*
□ *Trypanosoma brucei*

2. Filter targets based on:

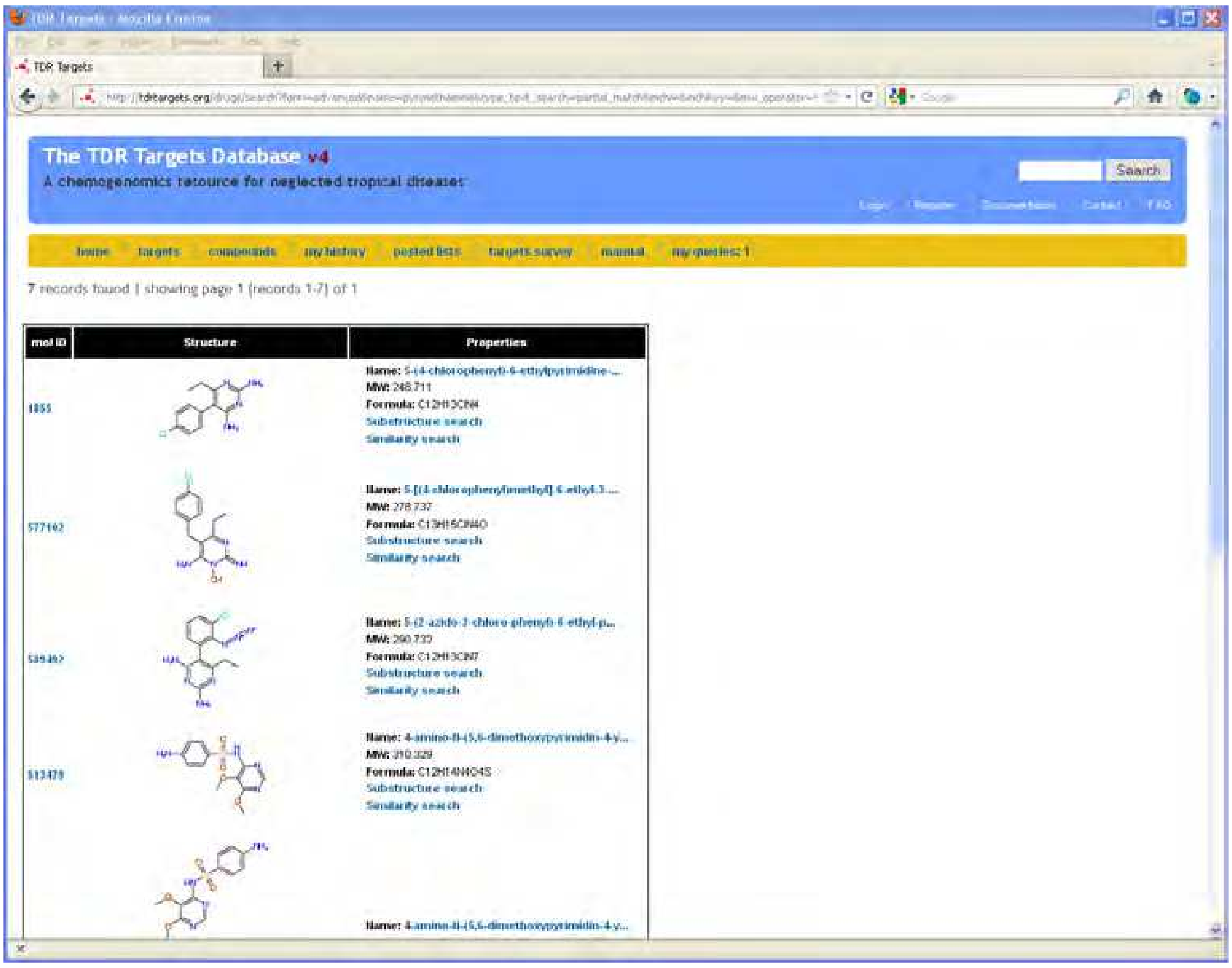
Name / Annotation

☑ Search for targets using keywords (names, functions, identifiers).

Name	kinase	[e.g. farnesyl, kinase, pyrophosphatase]
Identifier/Accession		[e.g. LmjF22.1360, PF11_0295, tb03370]
EC number		[e.g. 2.5.1.10, or use '*' or type 'any']
Gene Ontology		[GO id/term (GO:0020011, apicoplast)]
Pfam / Interpro domains		[accession number or description]

Fig. 13. The target search interface of the TDR Targets database. The different organisms available for searching are shown at the top, followed by possible search parameters at the bottom.

The target search functionality allows users to select proteins from the pathogen genome based on the very extensive criteria mentioned above, and queries may be saved for future reference. Searches are performed by first selecting the pathogen of interest, and this is followed by the addition of filtering criteria (Figure 13). Multiple queries may be executed, and subsquently cumulative search results may be generated by specifying unions, intersects or subtractions of the results. Weights and names may be associated with the different queries. When a target is explored, the view provides extensive information in all the categories listed above, sorted by category. Compound searches may be performed using text- or structure-based queries. Additionally, compound searches may be done based on chemical properties, activities and associated genes. A summary of resulting chemical structures is displayed together with basic properties, and a highly-detailed report for each compound may be obtained, including putative protein targets, activities and external resources for the compound (Figure 14). The TDR resource further provides a posted list feature for the sharing of data within the scientific community, as well as a targets survey page for gathering curated information from community experts.


TDR Targets
http://tdrtargets.org/drugs/search
The TDR Targets Database v4
A chemogenomics resource for neglected tropical diseases
Search
home
targets
compounds
my history
posted lists
targets survey
manual
my queries: 1

7 records found | showing page 1 (records 1-7) of 1

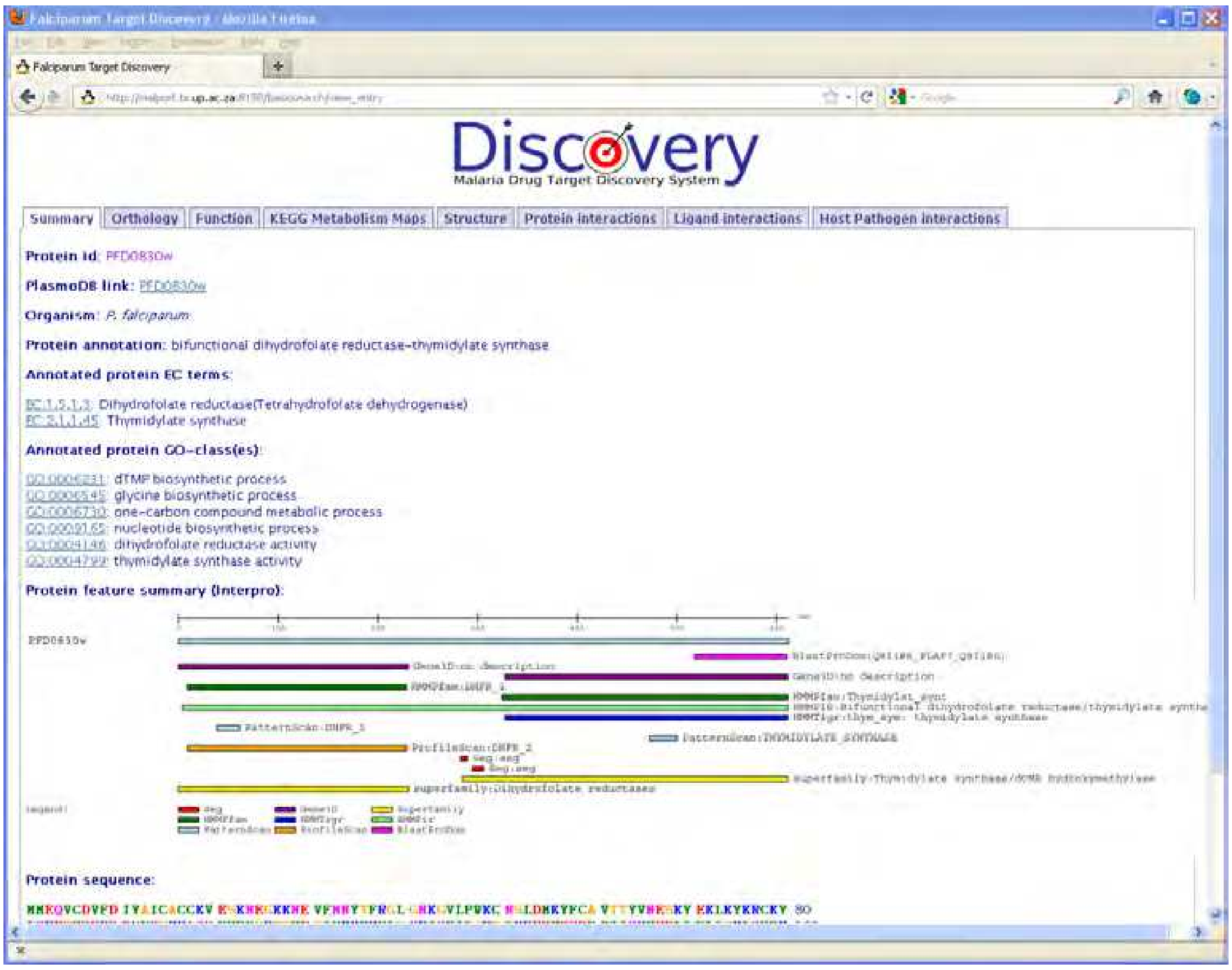
mol ID	Structure	Properties
1855		Name: 5-(4-chlorophenyl)-6-ethylpyrimidine-... MW: 248.711 Formula: C12H13ClN4 Substructure search Similarity search
577102		Name: 5-[(4-chlorophenyl)methyl]-6-ethyl-3-... MW: 278.737 Formula: C13H15ClN4O Substructure search Similarity search
535487		Name: 5-(2-azido-3-chloro-phenyl)-6-ethyl-p... MW: 290.733 Formula: C12H13ClN7 Substructure search Similarity search
513479		Name: 4-amino-N-(5,6-dimethoxypyrimidin-4-y... MW: 310.329 Formula: C12H14N4O4S Substructure search Similarity search
		Name: 4-amino-N-(5,6-dimethoxypyrimidin-4-y...

Fig. 14. The compound search results display of the TDR Targets database. Chemical identifiers and structures are shown on the left, with more detailed compound information on the right.

9.3 Discovery

URL: http://discovery.bi.up.ac.za

The Discovery resource is aimed at scientists who would like to explore primarily the *P. falciparum* genome for the selection of drug target proteins and lead compounds (Joubert *et al.*, 2009). Other parasite genomes included are *P. vivax*, *P. vivax*, *P. bergei*, *P. chabaudi* and *P. yoelii*. Also included for comparative purposes are the the human and mosquito genomes. The resource may be accessed by querying proteins or ligands. Proteins may be searched using simple keywords, combined terms or accession numbers. Protein data displayed includes orthology and sequence comparisons, ontology terms, functional annotations, metabolic pathways, structural information and possible ligand interactions based on orthology to sequences from DrugBank, PDB Ligand and KEGG. The new version of Discovery adds extensive data from ChEMBL, as well as literature mining and user annotation functionality (Figure 15).



Protein id: PFD0830w

PlasmoDB link: PFD0830w

Organism: *P. falciparum*

Protein annotation: bifunctional dihydrofolate reductase-thymidylate synthase

Annotated protein EC terms:

EC:1.5.1.3: Dihydrofolate reductase(Tetrahydrofolate dehydrogenase)
EC:2.1.1.45: Thymidylate synthase

Annotated protein GO-class(es):

GO:0006231: dTMP biosynthetic process
GO:0006545: glycine biosynthetic process
GO:0006730: one-carbon compound metabolic process
GO:0009165: nucleotide biosynthetic process
GO:0004146: dihydrofolate reductase activity
GO:0004799: thymidylate synthase activity

Protein feature summary (Interpro):

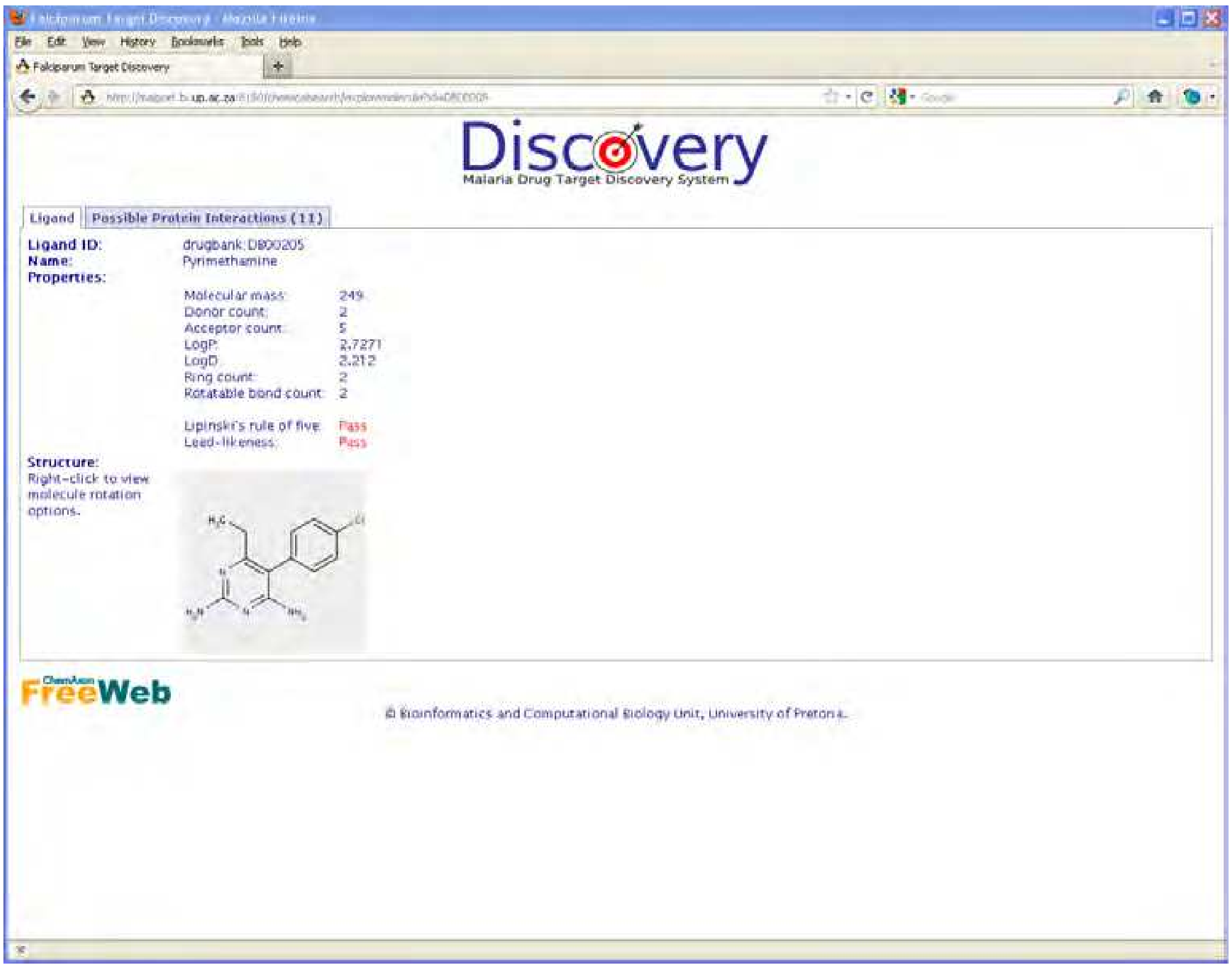
PFD0830w
BlastProDom
Gene3D:no description
HMMPfam:DHFR_1
HMMPfam:Thymidylat_synt
HMMPIR:Bifunctional dihydrofolate reductase/thymidylate synthe
HMMTigr:thye_sye: thymidylate synthase
PatternScan:DHFR_1
PatternScan:THYMIDYLATE_SYNTHASE
ProfileScan:DHFR_2
Seg:seg
superfamily:Thymidylate synthase/dCMP hydroxymethylase
superfamily:Dihydrofolate reductases
legend:
Seg
HMMPfam
PatternScan
Gene3D
HMMTigr
ProfileScan
Superfamily
HMMPir
BlastProDom

Protein sequence:

MMEQVCDVFD IYAICACCKV ESKNEGKKNE VFNNYTFRGL GNKGVLPWKC NSLDMKYFCA VTTYVNESKY EKLKYKRCKY 80

Fig. 15. The Discovery protein results view for PFD0830w (bifunctional dihydrofolate reductase-thymidylate synthase from *P. falciparum*). A summary of protein annotations is shown at the top, followed by a graphical representation of sequence motifs in the molecule. The other types of results that may be viewed can be seen as tabs at the top of the screen.

Chemical compounds may be searched using text terms or chemical structures, and searches are powered using the ChemAxon JChemBase software (http://www.chemaxon.com). Several different types of searches are available. Results are displayed in a tabular fashion, and compounds may be selected for further inspection. Additional detail is then provided, including the structure, ADMET properties and putative protein interactions (Figure 16).


File
Edit
View
History
Bookmarks
Tools
Help
Falciparum Target Discovery
Discovery
Malaria Drug Target Discovery System
Ligand
Possible Protein Interactions (11)
ChemAxon
FreeWeb

Ligand ID: drugbank:DB00205

Name: Pyrimethamine

Properties:

Molecular mass:	249
Donor count:	2
Acceptor count:	5
LogP:	2.7271
LogD:	2.212
Ring count:	2
Rotatable bond count:	2
Lipinski's rule of five:	Pass
Lead-likeness:	Pass

Structure: Right-click to view molecule rotation options.

© Bioinformatics and Computational Biology Unit, University of Pretoria.

Fig. 16. The Discovery chemical results view for pyrimethamine. The chemical properties are shown at the top, followed by the structure of the compound at the bottom. The tab for viewing possible protein interactions is visible at the top of the page.

10. Conclusion

Researchers face many different challenges when embarking on the discovery of new drug targets. Here we provide an overview on valuable resources that will enable scientists to perform comprehensive searches and make numerous comparisons between targets prior to selection. In addition these tools can be used in the evaluation and data mining of existing targets, which can easily be overlooked using conventional methods and should therefore be used before committing expensive and time-consuming resources. Of primary concern is the properties that need to be taken into account when selecting putative drug proteins or lead compounds. The term "drugability" is widely used for describing the suitability of both proteins and compounds as targets and leads, but specific properties and parameters related to this can be difficult to formalize. Various examples of studies proposing drugability properties and possible scoring systems are available in the literature (Crowther *et al.*, 2010; Fauman *et al.*, 2011; Halgren, 2009; Hasan *et al.*, 2006; Nicola *et al.*, 2008; Schneider, 2004).

Whereas the resources discussed in the first part of the chapter mostly address resources with single specialized foci, the latter two resources have attempted to integrate the most relevant properties related to protein and compound suitability for drug design during the design of the search strategies. These resources make it possible for researchers to start out with a large number of proteins or compounds, and filter them sequentially based on a series of properties related to "drugability", therefore producing a smaller list of candidate molecules that may be explored further in greater detail using the more focussed sites, and through experimental approaches. These integrated resources make it possible for researchers to in effect start out with all proteins encoded by a genome or all compounds in a database, and rapidly decrease the potential candidate list to a manageable number using the rational application of filters to properties of their choice.

11. Glossary

A brief glossary of some terms that may not be familiar to all readers:

Bayesian:	A method employing Bayes' theorem on conditional probability
BLAST:	Basic Local Alignment Search Tool which is used for searching sequences vs. databases
EC number:	Enzyme Commission number defining the type of reaction catalyzed by an enzyme
Essentiality:	An indication of whether the organism can survive after an enzyme has been knocked out or inhibited
Homology model:	A model of a protein structure calculated by using the structure of another homologous protein as template
Lipinksi rule-of-five:	A rule of five parameters proposed by Lipinski and colleagues which is used predict the pharmacological or biological activity of a chemical compound
Ontology:	A biological ontology term or identifier forms part of the Gene Ontology (GO) Consortium classification which provides a vocabulary describing gene product characteristics
ORF:	An Open Reading Frame in a nucleic acid sequence which potentially encodes a peptide region
Orthologs:	Sequences which are related to a common evolutionary ancestor but have diverged due to speciation
Paralogs:	Sequences which are related to a common evolutionary ancestor but have diverged due to gene duplication
Proteome:	The complete set of proteins produced by an organism or tissue

12. References

Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, Vol. 25(17), pp. 3389-3402.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, Vol. 25(1), pp. 25-29.

Aurrecoechea, C.; Brestelli, J.; Brunk, B. P.; Dommer, J.; Fischer, S.; Gajria, B.; Gao, X.; Gingle, A.; Grant, G.; Harb, O. S.; Heiges, M.; Innamorato, F.; Iodice, J.; Kissinger, J. C.; Kraemer, E.; Li, W.; Miller, J. A.; Nayak, V.; Pennington, C.; Pinney, D. F.; Roos, D. S.; Ross, C.; Stoeckert, C. J.; Treatman, C., & Wang, H. (2009). PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res, Vol. 37(Database issue), pp. D539-D543.

Aurrecoechea, C.; Brestelli, J.; Brunk, B. P.; Fischer, S.; Gajria, B.; Gao, X.; Gingle, A.; Grant, G.; Harb, O. S.; Heiges, M.; Innamorato, F.; Iodice, J.; Kissinger, J. C.; Kraemer, E. T.; Li, W.; Miller, J. A.; Nayak, V.; Pennington, C.; Pinney, D. F.; Roos, D. S.; Ross, C.; Srinivasamoorthy, G.; Stoeckert, C. J.; Thibodeau, R.; Treatman, C., & Wang, H. (2010). EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res, Vol. 38(Database issue), pp. D415-D419.

Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L. L.; Studholme, D. J.; Yeats, C., & Eddy, S. R. (2004). The Pfam protein families database. Nucleic Acids Res, Vol. 32(Database issue), pp. D138-D141.

Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol, Vol. 112(3), pp. 535-542.

Bozdech, Z.; Llinás, M.; Pulliam, B. L.; Wong, E. D.; Zhu, J., & DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol, Vol. 1(1), pp. E5.

Bréhélin, L.; Dufayard, J.-F., & Gascuel, O. (2008). PlasmoDraft: a database of *Plasmodium falciparum* gene function predictions based on postgenomic data. BMC Bioinformatics, Vol. 9, pp. 440.

Caspi, R.; Foerster, H.; Fulcher, C. A.; Hopkinson, R.; Ingraham, J.; Kaipa, P.; Krummenacker, M.; Paley, S.; Pick, J.; Rhee, S. Y.; Tissier, C.; Zhang, P., & Karp, P. D. (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res, Vol. 34(Database issue), pp. D511-D516.

Crowther, G. J.; Shanmugam, D.; Carmona, S. J.; Doyle, M. A.; Hertz-Fowler, C.; Berriman, M.; Nwaka, S.; Ralph, S. A.; Roos, D. S.; Voorhis, W. C. V., & Agüero, F. (2010). Identification of attractive drug targets in neglected-disease pathogens using an *in silico* approach. PLoS Negl Trop Dis, Vol. 4(8), pp. e804.

Date, S. V., & Stoeckert, C. J. (2006). Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. Genome Res, Vol. 16(4), pp. 542-549.

Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M.-Y.; Pieper, U., & Sali, A. (2006). Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics, Vol. Chapter 5, pp. Unit 5.6.

Fauman, E. B.; Rai, B. K., & Huang, E. S. (2011). Structure-based druggability assessment-identifying suitable targets for small molecule therapeutics. Curr Opin Chem Biol, Vol. 15(4), pp. 463-468.

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

Ginsburg, H. (2009). Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium. Trends Parasitol, Vol. 25(1), pp. 37-43.

Halgren, T. A. (2009). Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model, Vol. 49(2), pp. 377-389.

Hasan, S.; Daugelat, S.; Rao, P. S. S., & Schreiber, M. (2006). Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. PLoS Comput Biol, Vol. 2(6), pp. e61.

Hunter, S.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R. D.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Laugraud, A. l.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Mulder, N.; Natale, D.; Orengo, C.; Quinn, A. F.; Selengut, J. D.; Sigrist, C. J. A.; Thimma, M.; Thomas, P. D.; Valentin, F.; Wilson, D.; Wu, C. H., & Yeats, C. (2009). InterPro: the integrative protein signature database. Nucleic Acids Res, Vol. 37(Database issue), pp. D211-D215.

Joubert, F.; Harrison, C. M.; Koegelenberg, R. J.; Odendaal, C. J., & de Beer, T. A. P. (2009). Discovery: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. Malar J, Vol. 8, pp. 178.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, Vol. 28(1), pp. 27-30.

Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res, Vol. 38(Database issue), pp. D355-D360.

Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M., & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res, Vol. 34(Database issue), pp. D354-D357.

Kolaskar, A. S., & Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett, Vol. 276(1-2), pp. 172-174.

Li, L.; Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res, Vol. 13(9), pp. 2178-2189.

Lipinski, C. A.; Lombardo, F.; Dominy, B. W., & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. [Review]. Adv Drug Deliv Rev, Vol. 46(1-3), pp. 3-26.

Llinás, M.; Bozdech, Z.; Wong, E. D.; Adai, A. T., & DeRisi, J. L. (2006). Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. Nucleic Acids Res, Vol. 34(4), pp. 1166-1173.

Nicola, G.; Smith, C. A., & Abagyan, R. (2008). New method for the assessment of all drug-like pockets across a structural genome. J Comput Biol, Vol. 15(3), pp. 231-240.

Ortí, L.; Carbajo, R. J.; Pieper, U.; Eswar, N.; Maurer, S. M.; Rai, A. K.; Taylor, G.; Todd, M. H.; Pineda-Lucena, A.; Sali, A., & Marti-Renom, M. A. (2009). A kernel for the Tropical Disease Initiative. Nat Biotechnol, Vol. 27(4), pp. 320-321.

Overington, J. (2009). ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. J Comput Aided Mol Des, Vol. 23(4), pp. 195-198.

Pieper, U.; Eswar, N.; Webb, B. M.; Eramian, D.; Kelly, L.; Barkan, D. T.; Carter, H.; Mankoo, P.; Karchin, R.; Marti-Renom, M. A.; Davis, F. P., & Sali, A. (2009). MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res, Vol. 37(Database issue), pp. D347-D354.

Rice, P.; Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet, Vol. 16(6), pp. 276-277.

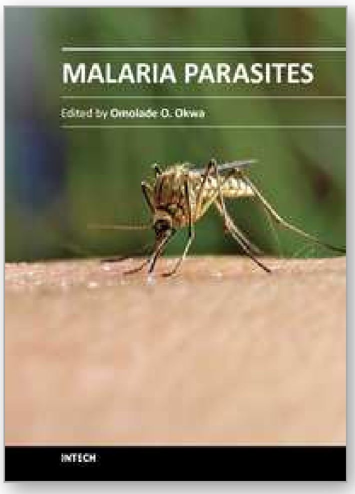
Scheer, M.; Grote, A.; Chang, A.; Schomburg, I.; Munaretto, C.; Rother, M.; Söhngen, C.; Stelzer, M.; Thiele, J., & Schomburg, D. (2011). BRENDA, the enzyme information system in 2011. Nucleic Acids Res, Vol. 39(Database issue), pp. D670-D676.

Schneider, M. (2004). A rational approach to maximize success rate in target discovery. Arch Pharm (Weinheim), Vol. 337(12), pp. 625-633.

Stein, L. D.; Mungall, C.; Shu, S.; Caudy, M.; Mangone, M.; Day, A.; Nickerson, E.; Stajich, J. E.; Harris, T. W.; Arva, A., & Lewis, S. (2002). The generic genome browser: a building block for a model organism system database. Genome Res, Vol. 12(10), pp. 1599-1610.

Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B., & Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids ResPDBLIgand, Vol. 36(Database issue), pp. D901-D906.

Yeh, I.; Hanekamp, T.; Tsoka, S.; Karp, P. D., & Altman, R. B. (2004). Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. Genome Res, Vol. 14(5), pp. 917-924.



Malaria Parasites

Edited by Dr. Omolade Okwa

ISBN 978-953-51-0326-4

Hard cover, 350 pages

Publisher InTech

Published online 30, March, 2012

Published in print edition March, 2012

Malaria is a global disease in the world today but most common in the poorest countries of the world, with 90% of deaths occurring in sub-Saharan Africa. This book provides information on global efforts made by scientist which cuts across the continents of the world. Concerted efforts such as symbiont based malaria control; new applications in avian malaria studies; development of humanized mice to study P.falciparium (the most virulent species of malaria parasite); and current issues in laboratory diagnosis will support the prompt treatment of malaria. Research is ultimately gaining more grounds in the quest to provide vaccine for the prevention of malaria. The book features research aimed to bring a lasting solution to the malaria problem and what we should be doing now to face malaria, which is definitely useful for health policies in the twenty first century.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821