MALARIA JOURNAL

METHODOLOGY Open Access

Discovery-2: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria

Phelelani T Mpangase[†], Michal J Szolkiewicz[†], Misha le Grange, Jeanré H Smit, Pieter B Burger and Fourie Joubert[*]

Abstract

Background: Drug resistance to anti-malarial compounds remains a serious problem, with resistance to newer pharmaceuticals developing at an alarming rate. The development of new anti-malarials remains a priority, and the rational selection of putative targets is a key element of this process. Discovery-2 is an update of the original Discovery *in silico* resource for the rational selection of putative drug target proteins, enabling researchers to obtain information for a protein which may be useful for the selection of putative drug targets, and to perform advanced filtering of proteins encoded by the malaria genome based on a series of molecular properties.

Methods: An updated *in silico* resource has been developed where researchers are able to mine information on malaria proteins and predicted ligands, as well as perform comparisons to the human and mosquito host characteristics. Protein properties used include: domains, motifs, EC numbers, GO terms, orthologs, protein-protein interactions, protein-ligand interactions. Newly added features include drugability measures from ChEMBL, automated literature relations and links to clinical trial information. Searching by chemical structure is also available.

Results: The updated functionality of the Discovery-2 resource is presented, together with a detailed case study of the *Plasmodium falciparum* S-adenosyl-L-homocysteine hydrolase (PfSAHH) protein. A short example of a chemical search with pyrimethamine is also illustrated.

Conclusion: The updated Discovery-2 resource allows researchers to obtain detailed properties of proteins from the malaria genome, which may be of interest in the target selection process, and to perform advanced filtering and selection of proteins based on a relevant range of molecular characteristics.

Background

Drug resistance to current anti-parasitic compounds has become widespread and is on the increase, including resistance to newer treatments such as artemisinin [1,2]. While in-depth studies are ongoing on a relatively small number of selected putative targets for future exploitation, not many resources are available that focus on performing data mining and target identification on the complete malaria genome, in concert with relations to chemical compounds. Currently available resources that may be useful for target identification include PlasmoDB [3], TDR Targets [4], PlasmoMap [5], the Tropical Diseases Kernel [6] and also the original version of Discovery [7].

Recent approaches have illustrated the value of predicting the association of chemical compounds with putative protein drug targets, especially when the targets of compounds such as the GSK dataset with known activity against the parasite may be extrapolated using protein-ligand interaction databases such as ChemProt [8,9]. The Discovery resource attempts to use a similar approach in associating chemical compounds with malaria proteins using sequence homology, and also selective chemical similarity searches. While the resource is focused primarily at *Plasmodium falciparum*, it contains information for all proteins from *Plasmodium vivax*, *Plasmodium yoelii*, *Plasmodium knowlesi*, *Plasmodium chabaudi* and *Plasmodium berghei* and also for the human and mosquito hosts. Protein information includes

[*] Correspondence: fjoubert@postino.up.ac.za
[†]Equal contributors
Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, Pretoria 0001, South Africa

sequences and annotations from PlasmoDB, Ensembl and VectorBase, functional predictions, gene ontology terms, orthology information, structural information, metabolic pathways, predicted putative protein-ligand interactions, druggability predictions, and literature links. The resource also contains chemical compounds from the ChEMBL database with chemical search functionality and putative ligand-protein prediction information. Protein searches may be performed using accession numbers, keywords or an advanced multi-parameter filtering interface. Chemical searches may be performed using keywords, SMILES strings or chemical structures.

This new implementation of Discovery (version 2) is a complete rewrite of the original Python-based system using Java and NetBeans with the implementation of automated updates based largely on web services. The previous version included information related to sequence features, orthology, ontology terms, structural information, metabolic pathways and protein ligand interactions. The new functionalities are the addition of expression information, literature information from PubMed abstracts, the implementation of druggability predictions from DrugEBIlity where available, the inclusion of malaria-related data from the clinical trials database and the implementation of Wiki-like user annotations. Specific advantages of Discovery-2 include extensive functionality to identify putative associations of proteins with ligands, an advanced chemical structure search interface encompassing the content of the ChEMBL database and an interactive system for refining putative target selections based on data mining of molecular properties. Additionally, the resource also contains data for the mosquito and human hosts, for easy comparative analysis.

Methods

Discovery-2 was developed in Java with NetBeans. The protein sequences for *P. berghei*, *P. chabaudi*, *P. falciparum*, *P. knowlesi*, *P. vivax* and *P. yoelii* were downloaded from PlasmoDB (currently version 8.1). The *Homo sapiens* and *Anopheles gambiae* proteins were downloaded from Ensembl (currently version 64) [10] and VectorBase (currently version 3.7) [11], respectively. UniProt accessions were assigned to the proteins using an identifier mapping file downloaded from UniProt [12]. The UniProt accessions were used to assign Gene Ontology (GO) annotations from UniProt-GOA database (currently October 2012) [13]. Each GO term was linked to the Gene Ontology database [14] and the AmiGO visualization tool [15] is used to view the GO terms [16]. To detect orthologous genes between the different species, ortholog clusters were generated with OrthoMCL using mostly default parameters with a percentage match cut-off of 50% [17]. Protein families, domains and functional sites were identified using InterProScan, which uses an ensemble of different methods to automatically annotate the function and structure of protein sequences [18]. To identify possible structures of related proteins, a sequence similarity search against the PDB database was carried out using NCBI BLAST, with an E-value cut-off value of 10^{-6}. Predicted 3D protein structures for *P. falciparum* and *Homo sapiens* were downloaded from the MODBASE database, which contains automatically generated homology models as part of the MODELER project [19]. Pathway information was obtained from KEGG PATHWAYS, MPMP (which is the primary malaria metabolic pathways site) [20] and Reactome [21]. Proteins were assigned pathway information by creating links to the pathways in the different databases they are involved in. The Enzyme Commission numbers were assigned to the proteins using enzyme data downloaded from the ENZYME database hosted by ExPASy [22]. For Plasmodium species, the EC numbers were obtained from PlasmoDB. The EC numbers were subsequently used to link proteins to the enzyme data in BRENDA [23] and KEGG ENZYME. Protein sequences from the DrugEBIlity database [24], which contains druggability predictions on protein structures from PDB, were kindly provided by the ChEMBL group. A sequence similarity search against the protein sequences from DrugEBIlity was carried out using NCBI BLAST, with an E-value cut-off of 10^{-6}, in order to assess the druggability of the matching proteins in Discovery-2 as well as closely-related proteins. The hits were linked to the DrugEBIlity database for more detailed information on the druggability calculations. Data for protein-ligand interactions was collected from the ChEMBL database. The predictions were generated by performing a BLAST search against the ChEMBL protein targets and finding matching functional domains that were generated through the InterProScan program. Expression data is from the study performed by DeRisi *et al.* [25]. The literature mining was done on PubMed's MEDLINE database abstracts (2012) by utilizing a form of the Aho-Corasick dictionary matching algorithm [26]. Articles are linked to proteins if they contain malaria-related keywords, and one of the protein's aliases. Additionally, abstracts are scanned for relational keywords, such as "interacts with", to positively or negatively connect proteins with chemical compounds from ChEMBL. Uncommon English words that occur frequently in articles linked to proteins are also stored. An advanced search function was designed with the help of the Hibernate ORM (an object-relation mapper to provide a Java object interface to the relational database), together with a custom-built criteria library, to ensure the fastest possible filtering of protein sequences. Filtering can be done on various fields including function, orthology and literature. Chemical searches are performed using the MarvinSketch Applet, and JChemBase (ChemAxon).

Results and discussion

A brief introduction to the resource

Accession and search functions

Multiple approaches can be used to access the resource in retrieving the desired information. Data on a specific protein of interest may be obtained through a basic or an advanced search. In the basic search, a user may enter a PlasmoDB/Ensembl identifier, a UniProt accession number or alternatively perform the search using a protein name. A built in auto-complete feature in the protein identifier and UniProt accession searches assists users as they enter characters in the search space. The advanced search function allows the users to filter proteins using different search criteria in a step-by-step manner to select proteins matching their requirements. The search criteria for filtering protein sequences include organism, protein name, function, gene ontology, the availability of structural information, orthology, protein-ligand interactions, and the availability of related PubMed articles. A list of proteins matching the search criteria is returned from which the user may select the desired protein to view detailed information. The information for each protein is split across tabs representing different categories of annotation data. These tabs include “Summary”, “Function”, “Gene Ontology”, “Orthology”, “Structure”, “Metabolic Pathways”, “Interactions”, “Expression”, “Protein-Ligand Interactions”, “Druggability” and “Literature”. An additional tab, “Comments”, is available where users may submit comments related to a protein in the resource.

Summary

The “Summary” tab displays the summary information for a protein together with its amino acid sequence. The protein names and synonyms are displayed in this tab together with the protein identifier used in the source database as well as the UniProt accession and KEGG GENE identifier if available.

Function

The “Function” tab displays the predicted protein families, domains, and functional sites predicted using InterProScan. A graphical representation of the InterPro features found in the protein sequence and a table describing the InterProScan analysis are shown.

Gene ontology

The “Gene ontology” tab displays the GO terms assigned to the protein. The GO terms are categorized according to the three domains, i.e. cellular component, molecular function and biological processes.

Orthology

The results from the OrthoMCL clustering are displayed in the “Orthology” tab. In this tab the multiple sequence alignment of the protein sequences that form the cluster, generated with T-Coffee, is displayed in the JalView applet along with detail of the degree of conservation of the proteins.

Structure

The “Structure” tab displays the results from a BLAST search against the PDB database, as well as structural predictions from MODBASE.

Metabolic pathway

The “Metabolic Pathways” tab displays MPMP, KEGG and Reactome pathways that the protein is involved in, as well as enzyme information if available.

Interactions

The “Interactions” tab displays the predicted protein-protein interactions between the query protein and other proteins. The interactions are grouped as obtained from each of the three interaction databases (IntAct, MINT and DIP).

Protein-ligand interactions

The “Protein-ligand interactions” tab displays the result of BLAST searches and domain matches against the CHEMBL targets database, and allows inspection of the chemicals that interact with the CHEMBL blast hits.

Duggability

The “Druggability” tab displays the results of BLAST searches against domains and proteins sequences in the DrugEBIlity database.

Expression

The “Expression” tab displays the expression of the selected protein over the 48 hour intraerythrocytic developmental cycle (IDC).

Literature

The “Literature” tab shows PubMed abstracts related to the protein of interest.

Chemical searches function

Chemical searches may be performed by drawing a structure in the MarvinSketch applet, loading an existing structure or by providing a SMILES string. Substructure, superstructure, exact and similarity searches are available. Results are displayed in a tabular fashion and include the structure, ADMET properties and links to possible interacting proteins.

Clinical trials

An additional new feature is the “Clinical Trials” page which summarizes clinical trials related to malaria that are registered on the clinicaltrials.gov website, which is an NIH-maintained resource that provides easy access to information on publicly and privately supported clinical studies on a wide range of diseases and conditions. This feature in Discovery-2 allows a user to initiate searches into Discovery using current/experimental drugs as starting point. Compounds under investigation may be sent to the chemical search module to find exact or similar compounds, and possible interacting proteins may be selected.

A sample investigation/Case study

To demonstrate the functionality available in Discovery-2, a sample investigation was performed on the enzyme S-adenosyl-L-homocysteine hydrolase (SAHH) from *P. falciparum.* PfSAHH is involved in cysteine and methionine metabolism, and has previously been proposed as a putative drug target [27]. The enzyme is responsible for the hydrolysis of S-adenosyl-L-homocysteine (SAH) to adenosine and L-homocysteine that is a reversible reduction/oxidation reaction involving nicotinamide adenine dinucleotide (NAD) as co-factor [27,28].

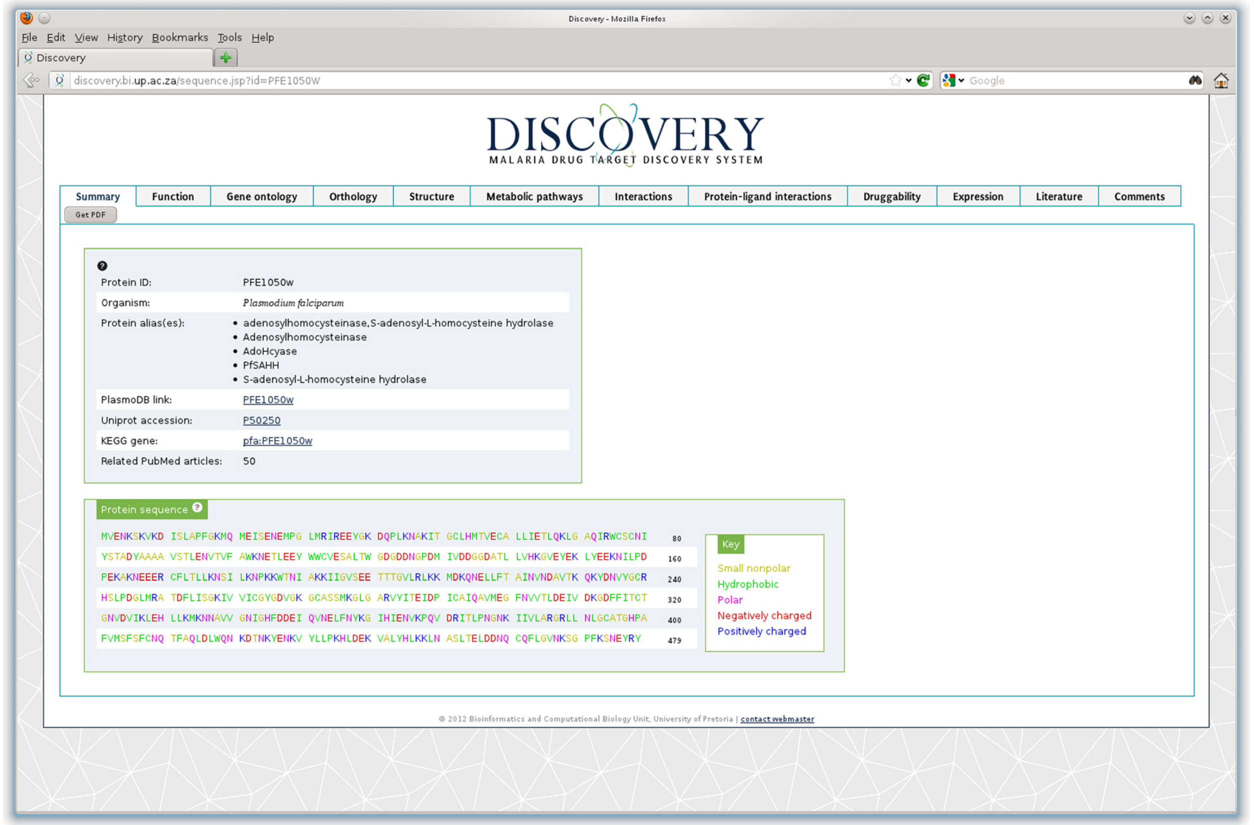
To search for PfSAHH in Discovery-2 using the basic search function, the protein entry could be retrieved by either using the PlasmoDB identifier PFE1050w, its UniProt accession P50250 or the enzyme name, S-adenosyl-L-homocysteine hydrolase.

Summary

The "Summary" tab (Figure 1) provides the user with the protein identifier from PlasmoDB, VectorBase or Ensembl depending on the organism from which it originates. In addition known aliases for PfSAHH such as "adenosylhomocysteinase" and "AdoHcyase" are given. Links are provided to the PlasmoDB, Uniprot, the KEGG database pertaining to PfSAHH. Finally, the number of associated papers and the amino acid sequence are given.

Function

Results from the "Function" tab (Figure 2) showed three different InterPro identifiers that matched the query 1) InterPro:IPR020082 matched S-adenosyl-L-homocysteine hydrolase, conserved site, 2) InterPro:IPR015878 (S-adenosyl-L-homocysteine hydrolase, NAD binding domain and 3) InterPro:IPR000043 that matched Adenosylhomocysteinase. Table 1 summarizes the InterPro entries matching the *Pf*SAHH protein sequence.


Discovery - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Discovery
discovery.bi.up.ac.za/sequence.jsp?id=PFE1050W
Google
DISCOVERY
MALARIA DRUG TARGET DISCOVERY SYSTEM
Summary
Function
Gene ontology
Orthology
Structure
Metabolic pathways
Interactions
Protein-ligand interactions
Druggability
Expression
Literature
Comments
Get PDF

Protein ID:	PFE1050w
Organism:	*Plasmodium falciparum*
Protein alias(es):	• adenosylhomocysteinase,S-adenosyl-L-homocysteine hydrolase • Adenosylhomocysteinase • AdoHcyase • PfSAHH • S-adenosyl-L-homocysteine hydrolase
PlasmoDB link:	PFE1050w
Uniprot accession:	P50250
KEGG gene:	pfa:PFE1050w
Related PubMed articles:	50

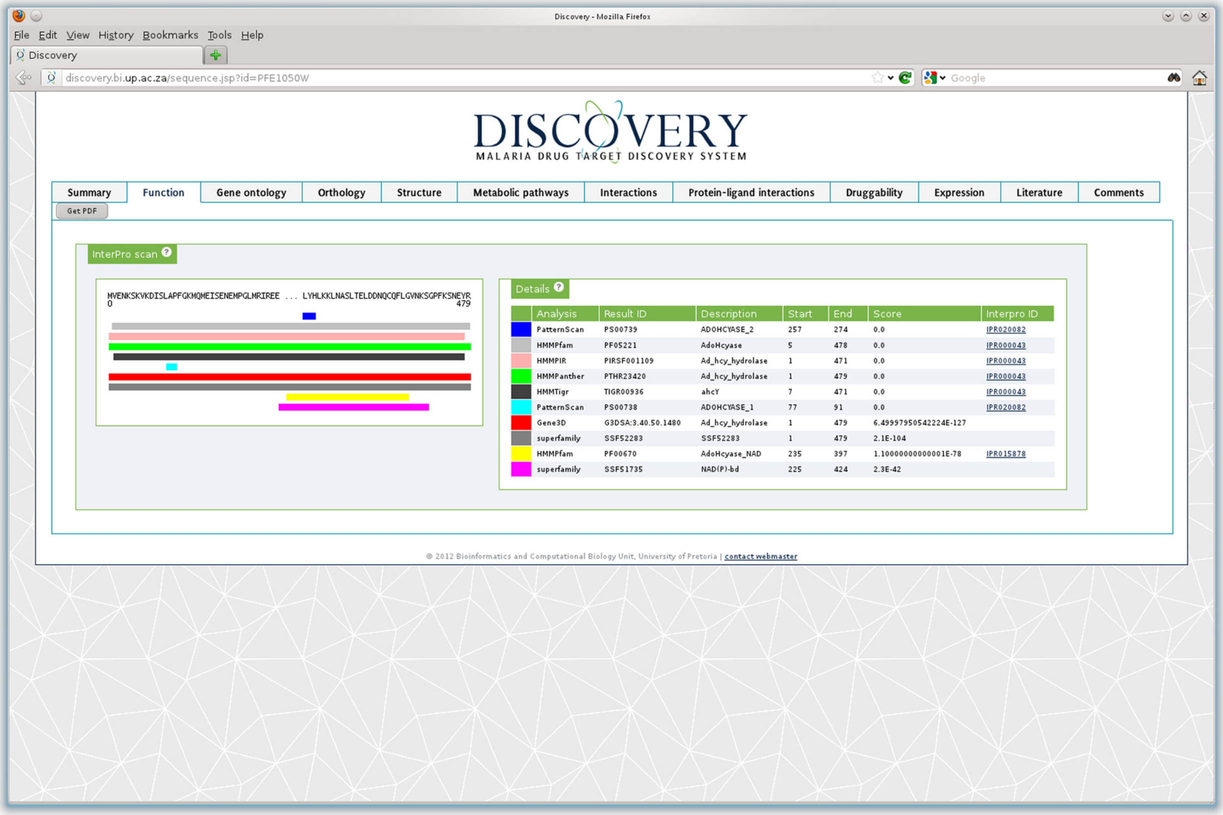
Protein sequence

```
MVENKSKVKD ISLAPFGKMQ MEISENEMPG LMRIREEYGK DQPLKNAKIT GCLHMTVECA LLIETLQKLG AQIRWCSCNI   80
YSTADYAAAA VSTLENVTVF AWKNETLEEY WWCVESALTW GDGDDNGPDM IVDDGGDATL LVHKGVEYEK LYEEKNILPD  160
PEKAKNEEER CFLTLLKNSI LKNPKKWTNI AKKIIGVSEE TTTGVLRLKK MDKQNELLFT AINVNDAVTK QKYDNVYGCR  240
HSLPDGLMRA TDFLISGKIV VICGYGDVGK GCASSMKGLG ARVYITEIDP ICAIQAVMEG FNVVTLDEIV DKGDFFITCT  320
GNVDVIKLEH LLKMKNNAVV GNIGHFDDEI QVNELFNYKG IHIENVKPQV DRITLPNGNK IIVLARGRLL NLGCATGHPA  400
FVMSFSFCNQ TFAQLDLWQN KDTNKYENKV YLLPKHLDEK VALYHLKKLN ASLTELDDNQ CQFLGVNKSG PFKSNEYRY   479
```

Key
Small nonpolar
Hydrophobic
Polar
Negatively charged
Positively charged

© 2012 Bioinformatics and Computational Biology Unit, University of Pretoria | contact webmaster

Figure 1 Summary view of the protein S-adenosyl-L-homocysteine hydrolase (PlasmoDB:PFE1050w), showing the basic protein information, links to other databases and protein sequence.


Discovery - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Discovery
discovery.bi.up.ac.za/sequence.jsp?id=PFE1050W
Google
DISCOVERY
MALARIA DRUG TARGET DISCOVERY SYSTEM
Summary
Function
Gene ontology
Orthology
Structure
Metabolic pathways
Interactions
Protein-ligand interactions
Druggability
Expression
Literature
Comments
Get PDF
InterPro scan
MVENKSKVKDISLAPFGKMQMEISENEMPGLMRIREE ... LYHLKKLNASLTELDDNQCQFLGVNKSGPFKSNEYR
0
479
Details
© 2012 Bioinformatics and Computational Biology Unit, University of Pretoria | contact webmaster

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

Figure 2 Functional prediction view of the protein S-adenosyl-L-homocysteine hydrolase (PlasmoDB:PFE1050w), showing a graphical overview as well as detail of domains and motifs found by InterProScan.

Gene ontology

Two GO terms, GO:0004013 (adenosylhomocysteinase activity) and GO:001-6787 (hydrolase activity) were found to be associated with PfSAHH at the molecular function level whereas GO:0006730 (one-carbon metabolic process) is associated with the protein at a biological process level. No GO terms were found to be associated with PfSAHH at cellular component level.

Orthology

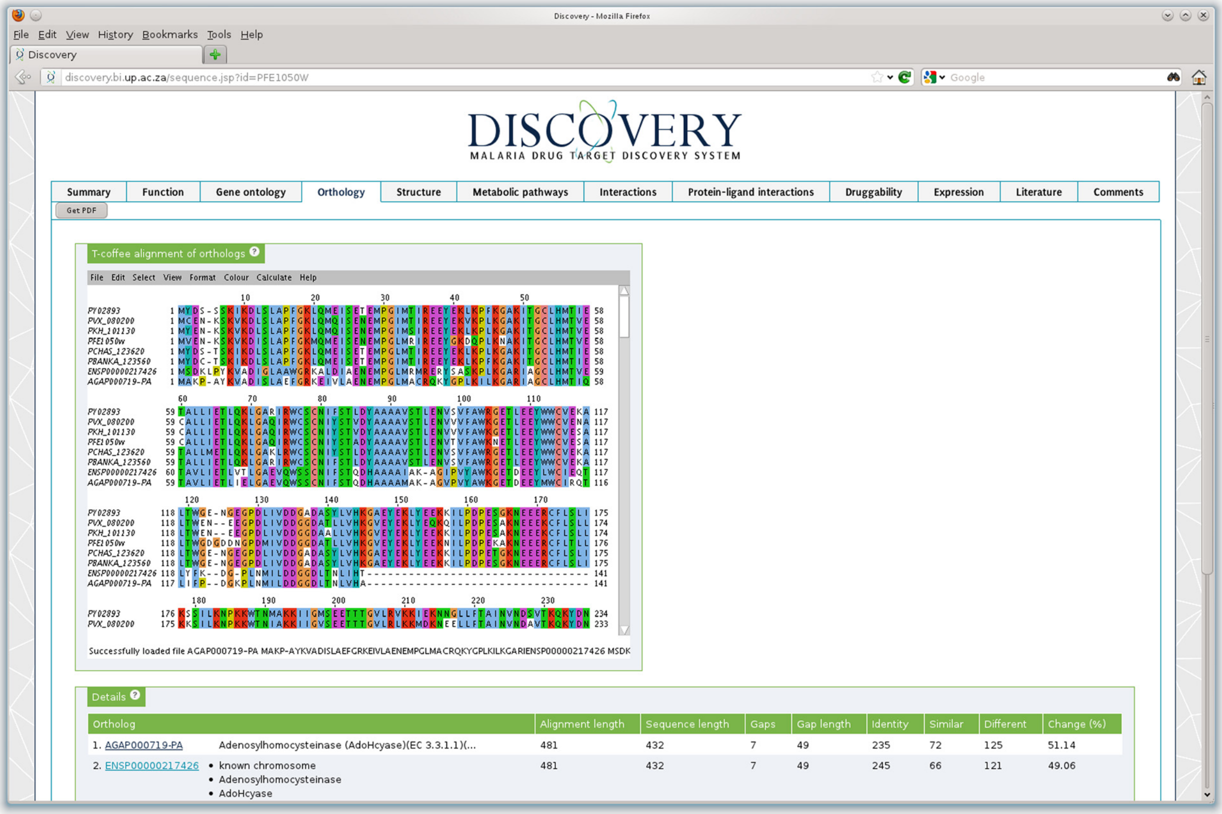
OrthoMCL clustering (Figure 3) showed SAHH to be present in all eight species studied within Discovery 2. From the Ortholog table it was noted that *A. gambiae*, and *H. sapiens* had the same alignment length to the Plasmodium species but differed in their sequence length. *A. gambiae*, and *H. sapiens* showed 7 gaps with a total gap length of 49 amino acids. From the T-coffee alignment an insert of 42 residues were identified in all the Plasmodium species starting at residue 145 in *P. falciparum*.

Structure

A BLAST search of the PfSAHH protein sequence against the PDB identified an exact protein sequence match for PfSAHH (PDB:1V8B). Related protein structures that have 70% coverage with the PfSAHH sequence are provided together with an alignment between the target and template. In addition two MODBASE models have been constructed for the PfSAHH protein, PFE1050w.1 and PFE1050w.2 which were built on templates PDB:1LI4 and PDB:1V8B, respectively.

Table 1 Summary of the InterPro signatures matching to *Pf*SAHH

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


Discovery - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Discovery
discovery.bi.up.ac.za/sequence.jsp?id=PFE1050W
Google
DISCOVERY
MALARIA DRUG TARGET DISCOVERY SYSTEM
Summary
Function
Gene ontology
Orthology
Structure
Metabolic pathways
Interactions
Protein-ligand interactions
Druggability
Expression
Literature
Comments
Get PDF
T-coffee alignment of orthologs
File Edit Select View Format Colour Calculate Help
Details

```
                          10        20        30        40        50
PY02893          1 MYDS-SSKIKDLSLAPFGKLQMEISETEMPGIMTIREEYEKLKPFKGAKITGCLHMTIE 58
PVX_080200       1 MCEN-KSKVKDLSLAPFGKLQMQISENEMPGIMTIREEYEKVKPLKGAKITGCLHMTVE 58
PKH_101130       1 MYEN-KSKVKDLSLAPFGKLQMQISENEMPGIMSIREEYEKLKPLKGAKITGCLHMTVE 58
PFE1050w         1 MVEN-KSKVKDISLAPFGKMQMEISENEMPGLMRIREEYGKDQPLKNAKITGCLHMTVE 58
PCHAS_123620     1 MYDS-TSKIKDLSLAPFGKLQMEISETEMPGLMTIREEYEKLKPLKGAKITGCLHMTIE 58
PBANKA_123560    1 MYDC-TSKIKDLSLAPFGKLQMEISETEMPGIMTIREEYEKLKPFKGAKITGCLHMTIE 58
ENSP00000217426  1 MSDKLPYKVADIGLAAWGRKALDIAENEMPGLMRMRERYSASKPLKGARIAGCLHMTVE 59
AGAP000719-PA    1 MAKP-AYKVADISLAEFGRKEIVLAENEMPGLMACRQKYGPLKILKGARIAGCLHMTIQ 58

                   60        70        80        90        100       110
PY02893         59 TALLIETLQKLGARIRWCSCNIFSTLDYAAAAVSTLENVSVFAWRGETLEEYWWCVEKA 117
PVX_080200      59 CALLIETLQKLGAQIRWCSCNIYSTVDYAAAAVSTLENVVVFAWKGETLEEYWWCVENA 117
PKH_101130      59 CALLIETLQKLGAQIRWCSCNIYSTVDYAAAAVSTLENVVVFAWKGETLEEYWWCVESA 117
PFE1050w        59 CALLIETLQKLGAQIRWCSCNIYSTADYAAAAVSTLENVTVFAWKNETLEEYWWCVESA 117
PCHAS_123620    59 TALLMETLQKLGAKLRWCSCNIYSTLDYAAAAVSTLENVSVFAWRGETLEEYWWCVEKA 117
PBANKA_123560   59 TALLIETLQKLGARIRWCSCNIFSTLDYAAAAVSTLENVSVFAWRGETLEEYWWCVEKA 117
ENSP00000217426 60 TAVLIETLVTLGAEVQWSSCNIFSTQDHAAAAIAK-AGIPVYAWKGETDEEYLWCIEQT 117
AGAP000719-PA   59 TAVLIETLIELGAEVQWSSCNIFSTQDHAAAAMAK-AGVPVYAWKGETDEEYMWCIRQT 116

                     120       130       140       150       160       170
PY02893         118 LTWGE-NGEGPDLIVDDGADASYLVHKGAEYEKLYEEKKILPDPESGKNEEERCFLSLI 175
PVX_080200      118 LTWEN--EEGPDLIVDDGGDATLLVHKGVEYEKLYEQKQILPDPESAKNEEEKCFLSLL 174
PKH_101130      118 LTWEN--EEGPDLIVDDGGDAALLVHKGVEYEKLYEEKKILPDPESAKNEEEKCFLSLL 174
PFE1050w        118 LTWGDGDDNGPDMIVDDGGDATLLVHKGVEYEKLYEEKNILPDPEKAKNEEERCFLTLL 176
PCHAS_123620    118 LTWGE-NGEGPDLIVDDGADASYLVHKGAEYEKLYEEKKILPDPETGKNEEERCFLSLI 175
PBANKA_123560   118 LTWGE-NGEGPDLIVDDGADASYLVHKGAEYEKLYEEKKILPDPESGKNEEERCFLSLI 175
ENSP00000217426 118 LYFK--DG-PLNMILDDGGDLTNLIHT-------------------------------- 141
AGAP000719-PA   117 LIFP--DGKPLNMILDDGGDLTNLVHA-------------------------------- 141

                      180       190       200       210       220       230
PY02893         176 KSSILKNPKKWTNMAKKIIGMSEETTTGVLRVKKIEKNNGLLFTAINVNDSVTKQKYDN 234
PVX_080200      175 KKSILKNPKKWTNIAKKIIGVSEETTTGVLRLKKMDKNEELLFTAINVNDAVTKQKYDN 233

Successfully loaded file AGAP000719-PA MAKP-AYKVADISLAEFGRKEIVLAENEMPGLMACRQKYGPLKILKGARIENSP00000217426 MSDK
```

Ortholog		Alignment length	Sequence length	Gaps	Gap length	Identity	Similar	Different	Change (%)
1. AGAP000719-PA	Adenosylhomocysteinase (AdoHcyase)(EC 3.3.1.1)(...	481	432	7	49	235	72	125	51.14
2. ENSP00000217426	• known chromosome • Adenosylhomocysteinase • AdoHcyase	481	432	7	49	245	66	121	49.06

Figure 3 Orthology information for S-adenosyl-L-homocysteine hydrolase (PlasmoDB:PFE1050w), showing a T-Coffee multiple alignment of the orthologs predicted for SPfSAHH in the other malaria species as well as the human and mosquito hosts.

Metabolic pathways

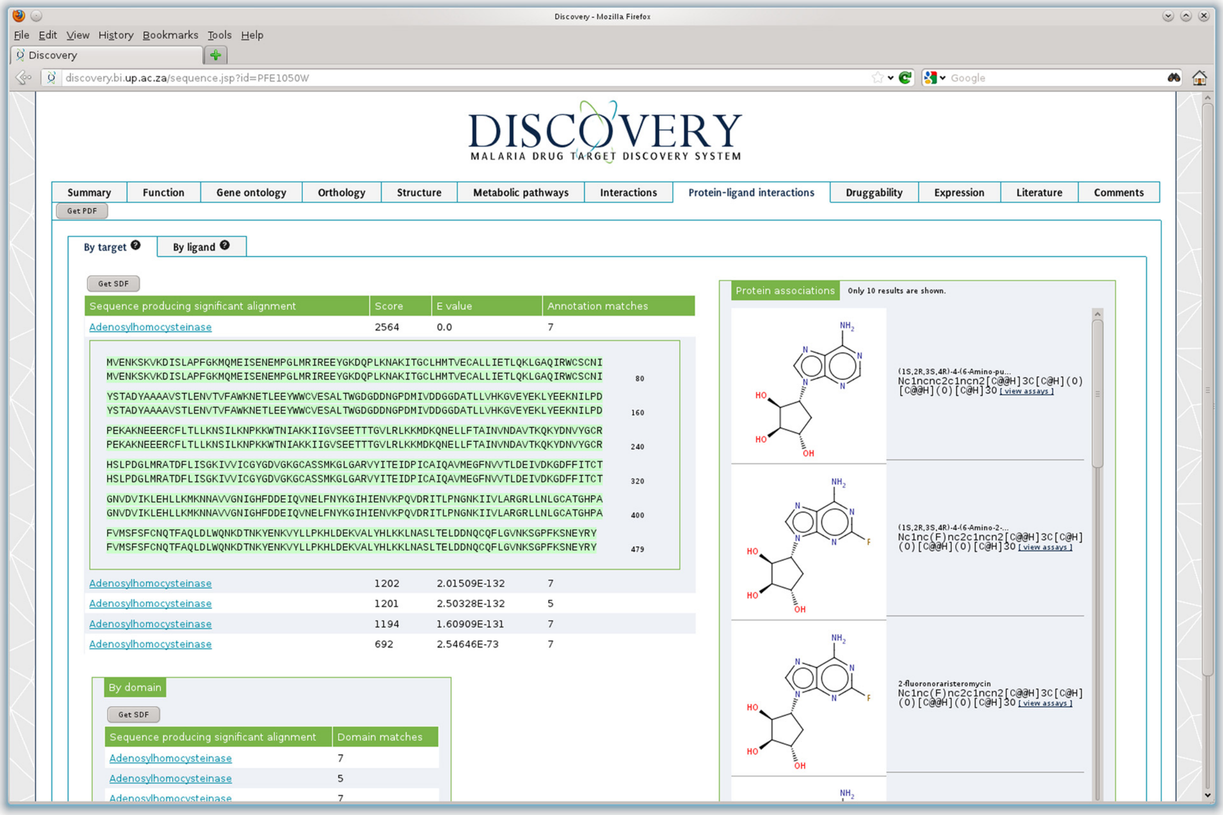
The "Metabolic Pathways" tab showed PfSAHH as adenosylhomocysteinase with its EC number, 3.3.1.1. The MPMP database showed the PfSAHH enzyme to be associated with methionine and polyamine metabolism, proteins targeted by the thioredoxin superfamily enzymes and S-glutathionylated proteins. The KEGG pathway associated the enzyme with cysteine and methionine metabolism and found it to be involved in various metabolic pathways (non-specific global map of all metabolic pathways) which include glycan biosynthesis, lipid, carbohydrate, amino acid, nucleotide, co-factor, vitamin, terpenoids, polyketide and energy metabolism. Reactome showed PfSAHH to be involved in metabolism of amino acids and derivatives as well as biological oxidations. The metabolic pathway function provides the user with an overview of the metabolic relevance of the target of interest.

Protein-protein interactions

The "Interactions" tab reports protein-protein interactions and had interaction data associated with PfSAHH within the IntAct and MINT databases. Nine interactions were associated with PfSAHH including itself. The UniProt accessions for these proteins that interact with the PfSAHH enzyme are UniProt:Q8I561 (conserved protein, unknown function), UniProt:Q8I2F7 (ring-exported protein [REX3]), UniProt:Q8IFP1 (U5 small nuclear ribonucleoprotein-specific protein, putative), UniProt: O96221 (Sec31p putative), UniProt:Q8IKB6 (histone deacetylase, putative), UniProt:Q8IIC8 (conserved protein, unknown function), UniProt:Q8IBL5 (conserved protein, unknown function), UniProt:Q8IAZ3 (Eukaryotic translation initiation factor 3 subunit G) and UniProt:Q8IJY1 (conserved protein, unknown function). All identified interactions were determined by the use of a two hybrid system that uses transcriptional activity as a measure of protein-protein interactions with all interactions therefore being demonstrated by physical associations.

Protein-ligand interactions

Protein-ligand interactions were identified by a BLAST search against the CHEMBL targets database, and were reported in the targets section with the ligands associated in the ligand section (Figure 4). PfSAHH was identified from the CHEMBL targets database along with SAHH from other organisms including *Mus musculus*, *Homo sapiens* and *Thermotoga maritima*, all showing


Discovery - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Discovery
discovery.bi.up.ac.za/sequence.jsp?id=PFE1050W
Google
DISCOVERY
MALARIA DRUG TARGET DISCOVERY SYSTEM
Summary
Function
Gene ontology
Orthology
Structure
Metabolic pathways
Interactions
Protein-ligand interactions
Druggability
Expression
Literature
Comments
Get PDF
By target
By ligand
Get SDF
Protein associations
Only 10 results are shown.
NH2
N
HO
OH
F

Sequence producing significant alignment	Score	E value	Annotation matches
Adenosylhomocysteinase	2564	0.0	7

```
MVENKSKVKDISLAPFGKMQMEISENEMPGLMRIREEYGKDQPLKNAKITGCLHMTVECALLIETLQKLGAQIRWCSCNI
MVENKSKVKDISLAPFGKMQMEISENEMPGLMRIREEYGKDQPLKNAKITGCLHMTVECALLIETLQKLGAQIRWCSCNI     80

YSTADYAAAAVSTLENVTVFAWKNETLEEYWWCVESALTWGDGDDNGPDMIVDDGGDATLLVHKGVEYEKLYEEKNILPD
YSTADYAAAAVSTLENVTVFAWKNETLEEYWWCVESALTWGDGDDNGPDMIVDDGGDATLLVHKGVEYEKLYEEKNILPD    160

PEKAKNEEERCFLTLLKNSILKNPKKWTNIAKKIIGVSEETTTGVLRLKKMDKQNELLFTAINVNDAVTKQKYDNVYGCR
PEKAKNEEERCFLTLLKNSILKNPKKWTNIAKKIIGVSEETTTGVLRLKKMDKQNELLFTAINVNDAVTKQKYDNVYGCR    240

HSLPDGLMRATDFLISGKIVVICGYGDVGKGCASSMKGLGARVYITEIDPICAIQAVMEGFNVVTLDEIVDKGDFFITCT
HSLPDGLMRATDFLISGKIVVICGYGDVGKGCASSMKGLGARVYITEIDPICAIQAVMEGFNVVTLDEIVDKGDFFITCT    320

GNVDVIKLEHLLKMKNNAVVGNIGHFDDEIQVNELFNYKGIHIENVKPQVDRITLPNGNKIIVLARGRLLNLGCATGHPA
GNVDVIKLEHLLKMKNNAVVGNIGHFDDEIQVNELFNYKGIHIENVKPQVDRITLPNGNKIIVLARGRLLNLGCATGHPA    400

FVMSFSFCNQTFAQLDLWQNKDTNKYENKVYLLPKHLDEKVALYHLKKLNASLTELDDNQCQFLGVNKSGPFKSNEYRY
FVMSFSFCNQTFAQLDLWQNKDTNKYENKVYLLPKHLDEKVALYHLKKLNASLTELDDNQCQFLGVNKSGPFKSNEYRY     479
```

Sequence producing significant alignment	Score	E value	Annotation matches
Adenosylhomocysteinase	1202	2.01509E-132	7
Adenosylhomocysteinase	1201	2.50328E-132	5
Adenosylhomocysteinase	1194	1.60909E-131	7
Adenosylhomocysteinase	692	2.54646E-73	7

By domain

Get SDF

Sequence producing significant alignment	Domain matches
Adenosylhomocysteinase	7
Adenosylhomocysteinase	5
Adenosylhomocysteinase	7

(1S,2R,3S,4R)-4-(6-Amino-pu...
Nc1ncnc2c1ncn2[C@@H]3C[C@H](O)[C@@H](O)[C@H]3O [view assays]

(1S,2R,3S,4R)-4-(6-Amino-2-...
Nc1nc(F)nc2c1ncn2[C@@H]3C[C@H](O)[C@@H](O)[C@H]3O [view assays]

2-fluoronoraristeromycin
Nc1nc(F)nc2c1ncn2[C@@H]3C[C@H](O)[C@@H](O)[C@H]3O [view assays]

Figure 4 Protein-ligand interaction view for S-adenosyl-L-homocysteine hydrolase (PlasmoDB:PFE1050w), showing predicted interactions with a series of ligands, based on a BLAST hit to ChEMBL. An alignment to the BLAST hit is also shown.

good BLAST hits. Domain matching of PfSAHH showed no additional matches compared to the BLAST results for the complete protein sequence. Two hundred and ninety-one chemical compounds were identified which have been tested for activity against PfSAHH and its homologs, 819 respective bioactivities involving those compounds were found.

Druggability

Druggability data for PfSAHH, showed that the molecules with the most significant BLAST domain matches from DrugEBIlity (PfSAHH crystal structure and PDB:1V8B) are not predicted to be druggable, although tractability is indicated. From the BLAST alignments PfSAHH (UniProt: P50250) had 0% overall druggability whilst the top two homologues UniProt:P60176 (SAHH from Mycobacterium tuberculosis) and UniProt:Q3JY79 (SAHH from *Burkholderia pseudomallei*, strain 1710b) had druggability scores of 48% and 45%, respectively.

Expression

The results in the expression view indicate that the protein is expressed predominantly in the trophozoite stage and is under-expressed in the early ring stage, indicating that it may have a role in parasite growth and preparation for division.

Literature

The literature functionality showed 26 articles related to SAHH from PubMed abstracts. The compounds cysteine, adenosine, aristeromycin, and ilimaquinone are shown as the main possible interactors with SAHH based on text relation mining of abstracts.

In summary, it was possible to demonstrate by means of a single search that the protein PfSAHH has ahcY, AdiHcyase and Ad_hcy_hydrolase signatures and that orthologs are found in the other Plasmodium species as well as the human and mosquito hosts. The Plasmodium enzymes have insertions in relation to the hosts' enzymes, and a crystal structure as well as homology models are available for the *P. falciparum* enzyme. Experimental protein-protein interaction evidence is available, and despite a relatively weak druggability score, a wide range of compounds that potentially interact with the enzyme are highlighted.

A sample investigation using a chemical structure

As a brief example of a compound-based search using a known anti-malarial drug, the SMILES string for pyrimethamine was provided, and an exact search was performed. The compound 5-(4-chlorophenyl)-6-ethylpyrimidine-2,4-diamine (pyrimethamine) was retrieved as the only hit. The compound information view provided detail about the molecular mass, hydrogen bond donor count, acceptor count, logP, logD, ring count and rotatable bond count, which are all relevant properties for the selection of druggable compounds. The compound passed the Lipinski rule of five [29]. When a search for putative targets was performed, a series of dihydrofolate reductase and bifunctional dihydrofolate reductase – thymidylate synthase enzymes were found in all the relevant species. This included the *Plasmodium falciparum* drug target, PFD0830w.

Conclusions

Discovery-2 is an update of the original Discovery resource, and is a complete rewrite of the system utilizing an updated database structure and new software technologies, and underlying functionality to perform automated updates using web services. A series of new features have been added such as relations to targets and compounds in ChEMBL, links to druggability information, automated literature links and relations to compounds in clinical trials.

The resource allows researchers to perform quick searches based on specific proteins, but to also perform mining and filtering of possible drug targets based on a series of molecular properties. Results provide the researcher with information about the protein's functional annotations, ontology terms, orthology to the different malaria parasite species and host proteins, structural information, pathway information, protein and ligand interaction information, druggability information and literature links. These result categories were selected to optimally assist researchers in selecting proteins for further detailed *in silico* and experimental investigation to assess the suitability of the protein as a putative drug target.

It is hoped that this resource enables the malaria community to quickly and easily obtain information related to the suitability of a protein as a drug target, and to perform the *in silico* selection of proteins from the malaria genome to effectively reduce the data space for further detailed experimentation.

Availability

Users can access the Discovery-2 resource at http://discovery.bi.up.ac.za.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FJ is the PI and grant holder. Core functionality was developed by JS. The protein-related functionality was developed by JS and PM. The chemical-related functionality was developed by MS and JS. Web-services and scoring was developed by JS, MLG and PB. All authors contributed to the writing of the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

We thank John Overington and the ChEMBL group for hosting PM and MS for a research visit. We are grateful to ChemAxon for the FreeWeb license. This project received funding from the National Research Foundation, The Medical Research Council and the University of Pretoria.

Received: 25 October 2012 Accepted: 18 March 2013
Published: 28 March 2013

References

1. O'Brien C, Henrich PP, Passi N, Fidock DA: **Recent clinical and molecular insights into emerging artemisinin resistance in *Plasmodium falciparum*.** *Curr Opin Infect Dis* 2011, **24:**570–577.
2. Dondorp AM, Nosten F, Yi P, Das D, Phyo AP, Tarning J, Lwin KM, Ariey F, Hanpithakpong W, Lee SJ, Ringwald P, Silamut K, Imwong M, Chotivanich K, Lim P, Herdman T, An SS, Yeung S, Singhasivanon P, Day NP, Lindegardh N, Socheat D, White NJ: **Artemisinin resistance in *Plasmodium falciparum* malaria.** *N Engl J Med* 2009, **361:**455–467.
3. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ Jr, Treatman C, Wang H: **PlasmoDB: a functional genomic database for malaria parasites.** *Nucleic Acids Res* 2009, **37:**D539–D543.
4. Aguero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, Carmona S, Carruthers IM, Chan AW, Chen F, Crowther GJ, Doyle MA, Hertz-Fowler C, Hopkins AL, McAllister G, Nwaka S, Overington JP, Pain A, Paolini GV, Pieper U, Ralph SA, Riechers A, Roos DS, Sali A, Shanmugam D, Suzuki T, Van Voorhis WC, Verlinde CL: **Genomic-scale prioritization of drug targets: the TDR Targets database.** *Nat Rev Drug Discov* 2008, **7:**900–907.
5. Date SV, Stoeckert CJ Jr: **Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale.** *Genome Res* 2006, **16:**542–549.
6. Orti L, Carbajo RJ, Pieper U, Eswar N, Maurer SM, Rai AK, Taylor G, Todd MH, Pineda-Lucena A, Sali A, Marti-Renom MA: **A kernel for open source drug discovery in tropical diseases.** *PLoS Negl Trop Dis* 2009, **3:**e418.
7. Joubert F, Harrison CM, Koegelenberg RJ, Odendaal CJ, de Beer TA: **Discovery: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria.** *Malar J* 2009, **8:**178.
8. Jensen K, Plichta D, Panagiotou G, Kouskoumvekaki I: **Mapping the genome of *Plasmodium falciparum* on the drug-like chemical space reveals novel anti-malarial targets and potential drug leads.** *Mol Biosyst* 2012, **8:**1678–1685.
9. Taboureau O, Nielsen SK, Audouze K, Weinhold N, Edsgard D, Roque FS, Kouskoumvekaki I, Bora A, Curpan R, Jensen TS, Brunak S, Oprea TI: **ChemProt: a disease chemical biology database.** *Nucleic Acids Res* 2011, **39:**D367–D372.
10. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM: **Ensembl 2012.** *Nucleic Acids Res* 2012, **40:**D84–D90.
11. Megy K, Emrich SJ, Lawson D, Campbell D, Dialynas E, Hughes DS, Koscielny G, Louis C, Maccallum RM, Redmond SN, Sheehan A, Topalis P, Wilson D: **VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics.** *Nucleic Acids Res* 2012, **40:**D729–D734.
12. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data.** *Database (Oxford)* 2011, **2011:**bar009.
13. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, Bely B, Browne P, Mun Chan W, Eberhardt R, Gardner M, Laiho K, Legge

D, Magrane M, Pichler K, Poggioli D, Sehra H, Auchincloss A, Axelsen K, Blatter MC, Boutet E, Braconi-Quintaje S, Breuza L, Bridge A, Coudert E, Estreicher A, Famiglietti L, Ferro-Rojas S, Feuermann M, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jimenez S, Jungo F, Keller G, Lemercier P, Lieberherr D, Masson P, Moinat M, Pedruzzi I, Poux S, Rivoire C, Roechert B, Schneider M, Stutz A, Sundaram S, Tognolli M, Bougueleret L, Argoud-Puy G, Cusin I, Duek-Roggli P, Xenarios I, Apweiler R: **The UniProt-GO Annotation database in 2011.** *Nucleic Acids Res* 2012, **40**:D565–D570.

14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–29.
15. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, **40**:D109–D114.
16. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25**:288–289.
17. Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ Jr: **Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups.** *Curr Protoc Bioinformatics* 2011, **Chapter 6**. Unit 6.12. 11–19.
18. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY: **InterPro in 2011: new developments in the family and domain prediction database.** *Nucleic Acids Res* 2012, **40**:D306–312.
19. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng EC, Pettersen EF, Huang CC, Datta RS, Sampathkumar P, Madhusudhan MS, Sjolander K, Ferrin TE, Burley SK, Sali A: **ModBase, a database of annotated comparative protein structure models, and associated resources.** *Nucleic Acids Res* 2011, **39**:D465–474.
20. Ginsburg H, Tilley L: ***Plasmodium falciparum*** **metabolic pathways (MPMP) project upgraded with a database of subcellular locations of gene products.** *Trends Parasitol* 2011, **27**:285–286.
21. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L: **Reactome: a database of reactions, pathways and biological processes.** *Nucleic Acids Res* 2011, **39**:D691–697.
22. Bairoch A: **The ENZYME database in 2000.** *Nucleic Acids Res* 2000, **28**:304–305.
23. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Sohngen C, Stelzer M, Thiele J, Schomburg D: **BRENDA, the enzyme information system in 2011.** *Nucleic Acids Res* 2011, 39:D670–676.
24. DrugEBIlity:. https://www.ebi.ac.uk/chembl/drugebility/.
25. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of** ***Plasmodium falciparum*****.** *PLoS Biol* 2003, **1**:E5.
26. Aho A, Corasick M: **Efficient string matching: An aid to bibliographic search.** *Comm ACM* 1975, **18**:333–240.
27. Shuto S, Minakawa N, Niizuma S, Kim HS, Wataya Y, Matsuda A: **New neplanocin analogues. 12. Alternative synthesis and antimalarial effect of (6'R)-6'-C-methylneplanocin A, a potent AdoHcy hydrolase inhibitor.** *J Med Chem* 2002, **45**:748–751.
28. Creedon KA, Rathod PK, Wellems TE: ***Plasmodium falciparum*** **S-adenosylhomocysteine hydrolase. cDNA identification, predicted protein sequence, and expression in** ***Escherichia coli*****.** *J Biol Chem* 1994, **269**:16364–16370.
29. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev* 1997, **23**:3–25.

doi:10.1186/1475-2875-12-116
Cite this article as: Mpangase *et al.*: Discovery-2: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. *Malaria Journal* 2013 **12**:116.