Journal of Microbiological Methods 109 (2015) 1–8

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

In silico comparative genomics analysis of *Plasmodium falciparum* for the identification of putative essential genes and therapeutic candidates

Subhashree Rout [a], David Charles Warhurst [b], Mrutyunjay Suar [a], Rajani Kanta Mahapatra [a],*

[a] School of Biotechnology, KIIT University, Bhubaneswar 751024, Orissa, India
[b] Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom

ARTICLE INFO

Article history:
Received 3 November 2014
Received in revised form 27 November 2014
Accepted 27 November 2014
Available online 5 December 2014

Keywords:
P. falciparum
Comparative genomics
Drug targets
Molecular docking

ABSTRACT

A sequence of computational methods was used for predicting novel drug targets against drug resistant malaria parasite *Plasmodium falciparum.* Comparative genomics, orthologous protein analysis among same and other malaria parasites and protein–protein interaction study provide us new insights into determining the essential genes and novel therapeutic candidates. Among the predicted list of 21 essential proteins from unique pathways, 11 proteins were prioritized as anti-malarial drug targets. As a case study, we built homology models of two uncharacterized proteins using MODELLER v9.13 software from possible templates. Functional annotation of these proteins was done by the InterPro databases and from ProBiS server by comparison of predicted binding site residues. The model has been subjected to *in silico* docking study with screened potent lead compounds from the ZINC database by Dock Blaster software using AutoDock 4. Results from this study facilitate the selection of proteins and putative inhibitors for entry into drug design production pipelines.

1. Introduction

Malaria, the widespread tropical parasitic disease, needs new antimalarial drugs and vaccines urgently, particularly to prevent its deadly effects seen mostly in children and during pregnancy. In 2013, 97 countries had ongoing malaria transmission with an estimated 3.4 billion people currently at risk of malaria (World Health Organization, 2013). Five *Plasmodium* species, namely, *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*, and *Plasmodium knowlesi* cause malaria in humans (Arama and Blomberg, 2014) among which *P. falciparum* is responsible for most morbidity and mortality (Miller et al., 2013). Current malaria chemotherapies are subject to resistance, and now even the artemisinins are seen as possibly a fading hope (Andrews et al., 2014). There is a continuous need to search for additional drug targets for better protection and long term effectiveness.

This study employs computational approaches for finding suitable antimalarial drug targets through comparative metabolic pathway analysis of pathogen and host. Essentiality of proteins of interest which are non-homologous to the human host can be predicted if the protein is found in falciparum and other malaria parasite proteomes (Ludin et al., 2012) and has a high functional association with other proteins there through protein–protein interaction network(s) (Kushwaha and Shakya, 2010). Despite the global importance of *P. falciparum*, most components of the pathogen proteome have not yet been characterized experimentally. In the present study we have made an attempt to determine the structure and functions of some uncharacterized proteins computationally, as suitable drug targets.

2. Methods

A systematic workflow was defined that involved several bioinformatics tools, databases and drug target prioritization parameters (Fig. 1), with the goal of obtaining information about the drug targets in the *P. falciparum* genome but absent in its host, therefore avoiding any potential side effects.

2.1. Identification of metabolic pathways of pathogen and host

Metabolic pathway information of *P. falciparum* 3D7 and *Homo sapiens* was taken from the PlasmoDB (Yeh et al., 2004) and KEGG pathway databases (Kanehisa et al., 2010) respectively. Manual comparisons were made between pathogen and host pathways. Pathways which appeared in both host and pathogen were considered as common and those which did not were considered as unique in nature.

2.2. Identification of non-homologous proteins

The corresponding protein sequences from unique pathways of pathogen were taken from the Uniprot database (Boeckmann et al., 2003) with reference to Uniprot accession number from PlasmoDB. They were subjected to a BLASTP search (Altschul et al., 1997) against

* Corresponding author. Tel.: +91 9337627961.
E-mail address: rmahapatra@kiitbiotech.ac.in (R.K. Mahapatra).

http://dx.doi.org/10.1016/j.mimet.2014.11.016
0167-7012/© 2014 Elsevier B.V. All rights reserved.

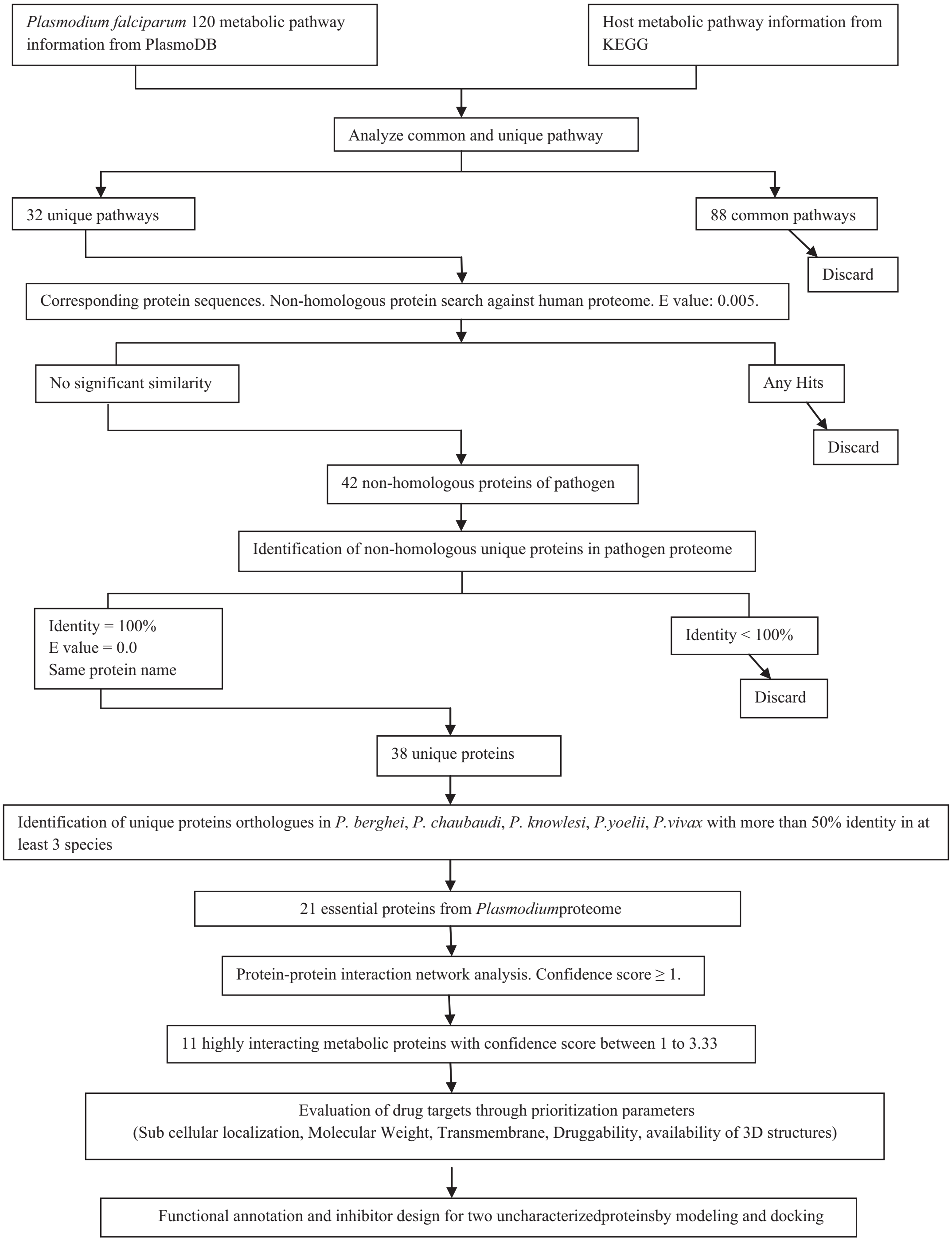
Plasmodium falciparum 120 metabolic pathway information from PlasmoDB
Host metabolic pathway information from KEGG
Analyze common and unique pathway
32 unique pathways
88 common pathways
Discard
Corresponding protein sequences. Non-homologous protein search against human proteome. E value: 0.005.
No significant similarity
Any Hits
Discard
42 non-homologous proteins of pathogen
Identification of non-homologous unique proteins in pathogen proteome
Identity = 100%
E value = 0.0
Same protein name
Identity < 100%
Discard
38 unique proteins
Identification of unique proteins orthologues in *P. berghei*, *P. chaubaudi*, *P. knowlesi*, *P.yoelii*, *P.vivax* with more than 50% identity in at least 3 species
21 essential proteins from *Plasmodium*proteome
Protein-protein interaction network analysis. Confidence score ≥ 1.
11 highly interacting metabolic proteins with confidence score between 1 to 3.33
Evaluation of drug targets through prioritization parameters
(Sub cellular localization, Molecular Weight, Transmembrane, Druggability, availability of 3D structures)
Functional annotation and inhibitor design for two uncharacterizedproteinsby modeling and docking

Fig. 1. Schematic representation of steps involved in drug target identification of *P. falciparum* through computational methods.

the human proteome with the e-value threshold set to 0.005 (Anishetty et al., 2005; Ghosh et al., 2014; Chawley et al., 2014). We adopted a rigorous way of determining no significant similarity for non-homologous proteins after applying filters.

2.3. Unique and conserved (essential) protein prediction in *P. falciparum*

Essentiality can be further predicted by identifying unique proteins in *P. falciparum* and other *Plasmodium* proteomes (Ludin et al., 2012). Non-homologous protein sequences from unique pathways were subjected to a manual BLASTP search for no other match in *P. falciparum* proteome with a default e-value threshold. These proteins are regarded as unique proteins in the *P. falciparum* genome. We used an OrthoMCL search (Chen et al., 2006) to identify unique proteins that are having the same orthologous group in other *Plasmodium* species. We considered the default e-value for putative orthologous based study. Additionally, a BLASTP search was employed for identification of conserved proteins against other *Plasmodium* proteomes namely *Plasmodium berghei*, *Plasmodium chabaudi*, *P. knowlesi*, *Plasmodium yoelii*, and *P. vivax* with an e-value threshold of $1e^{-5}$. We used a cutoff score of 50% homology in at least three *Plasmodium* species. This is a validation study searched by the OrthoMCL database for identification of conserved orthologous. These unique and orthologous proteins are proposed to be conserved and essential in *P. falciparum*.

2.4. Protein–protein interaction network analysis

Protein interaction network analysis was performed to find out the most likely metabolic functional associations using the protein interaction database STRING (Franceschini et al., 2003; von Mering et al., 2003) among the unique and conserved proteins of *P. falciparum*. In STRING, methods namely neighborhood, co-occurrence, gene fusion, experimental methods, co-expression and database information have been used for potential metabolic interaction identification (Kushwaha and Shakya, 2010). The interaction confidence score of the association is calculated for best associations as follows:

$$\text{Confidence score of the target} = \frac{\text{No. of interactants of target by used method}}{\text{Total no. of methods used}}.$$

2.5. Drug target prioritization

There are certain measures for determining suitable drug targets which involves molecular weight (MW) calculation using computational tools. Transmembrane domain (TMD) prediction was made using TMHMM server (Krogh et al., 2001). The PDB database (Bernstein et al., 1977) and Mod-Base (Berman et al., 2000; Pieper et al., 2011) were searched for solved 3D structures. ESLpred (Bhasin and Raghava, 2004), BaCelLo (Pierleoni et al., 2006) and EuLoc (Chang et al., 2013) servers were referred for sub-cellular localization to identify the cytoplasmic and membrane proteins for understanding its potential function and their role in biological processes.

2.6. Homology modeling and model optimization

A homology model of two uncharacterized proteins of *P. falciparum* was built using MODELLER v9.13 (Sali and Blundell, 1993). The models with the lowest DOPE scores generated by MODELLER v9.13 were selected for further study. The loop refinement tool provided by MODELLER v9.13 was implemented for loop refining of the modeled structures. The quality of refined models was measured using SAVES metaserver comprising of PROCHECK (Laskowski et al., 1993), VERIFY3D (Eisenberg et al., 1997), ERRAT (Colovos and Yeates, 1993) and PROVE servers (Pontius et al., 1996). Structure visualization and image production were carried out using Accelrys Discovery Studio 4.0 and PyMOL software. ProQ (Wallner and Elofsson, 2003) and ProSA (Wiederstein and Sippl, 2007) web servers are also used for analyzing the quality and native conformation of the generated 3D model. The probable function of these uncharacterized proteins can be determined from the Pfam (Finn et al., 2014) and InterPro databases (Mulder and Apweiler, 2008).

2.7. Docking

The ProBiS server (Konc et al., 2013) was used for identifying binding site residues of uncharacterized proteins. It is a searchable repository of pre-calculated local structural alignments in proteins detected by the ProBiS algorithm in the Protein Data Bank. Identification of functionally important binding regions of the protein is facilitated by structural similarity scores mapped to the query protein structure. The database also produces a possible Z-score and alignment score which helps in selecting binding site residues. Potent lead inhibitory compounds of both protein models were screened from the ZINC database (Irwin and Shoichet, 2005) by Dock Blaster (Irwin et al., 2009) software and used as inhibitors for their docking studies using AutoDock 4 (Morris et al., 2009).

3. Results and discussion

3.1. Identification of metabolic pathways and non-homologous proteins

Metabolic pathway information of *P. falciparum* 3D7 was taken from PlasmoDB, the official database of the malaria parasite genome project and contains complete genome information for *Plasmodium* spp. The metabolic pathway information of 120 different pathways was retrieved from PlasmoDB. PlasmoDB contains information of pathways and complete genomic sequences of the *Plasmodium* genus. Host metabolic pathway information was taken from the KEGG database. Pathogen and human host metabolic pathways were compared manually for common and unique pathways among them. 88 pathways were found to be common and 32 were unique between host and parasite. For our further studies unique pathways were considered. Aminobenzoate degradation, stilbenoid, diarylheptanoid and gingerol biosynthesis, methane metabolism, tetracycline biosynthesis, polycyclic aromatic hydrocarbon degradation, benzoxazinoid biosynthesis, insect hormone biosynthesis, flavonoid biosynthesis, carotenoid biosynthesis and bisphenol degradation are some unique pathways of the parasite. A total of 1040 proteins were reported from 32 unique pathways.

Non-homologous protein search was employed as our next step of analysis. Unique pathway proteins were subjected to a BLASTP search against the host proteome with the e-value threshold set to 0.005. Proteins showing “no significant similarity” were considered as non-homologous proteins against host. 42 proteins were identified as non-homologous proteins and subjected to further screening procedures for potential drug and vaccine targets. Some hypothetical proteins of *P. falciparum* were also reported through comparisons which can be considered as potential antimalarial drug targets. In this present study we tried to shed some light on some uncharacterized proteins of pathogen.

3.2. Essential proteins in P. falciparum

42 proteins were examined for its essentiality to parasite survival. As the DEG (Database of Essential Genes) did not provide much information about essential proteins in *Plasmodium*, here we employed another approach, of essential protein prediction based on search for unique proteins within *Plasmodium* species. Proteins can be essential if (i) no other match of proteins in *P. falciparum* proteome and (ii) those proteins are orthologues in other *Plasmodium* species. Non-homologous proteins were subjected to a BLASTP search against the *P. falciparum* proteome with default parameters. 38 proteins were filtered through the procedure and found to be unique to the pathogen (Table S1). We employed an OrthoMCL search for the prediction of orthologous proteins in other malaria parasites. OrthoMCL provides a scalable method for constructing orthologous groups using a Markov Cluster algorithm to group (putative) orthologs. Furthermore, these orthologous proteins were subjected to a BLASTP search against five other causative agents of malaria namely *P. vivax*, *P. yoelii*, *P. berghei*, *P. knowlesi* and *P. chabaudi* proteomes with an e-value of $1e^{-5}$. Unique orthologous proteins were selected based on identity ≥50% in at least 3 other *Plasmodium* species (Table S2). As these proteins are unique to *P. falciparum* and orthologous to other *Plasmodium* species, they can be considered as conserved essential proteins. 21 out of 38 proteins were filtered to be conserved and essential in nature.

3.3. Protein–protein interaction network analysis and drug target prioritization

21 unique proteins were mapped for the analysis of functional associations among them using the STRING database (Fig. 2). This protein–protein interaction network displays clear associations of proteins. The interaction view provides a valuable framework for a better understanding of the functional organization of the proteome.

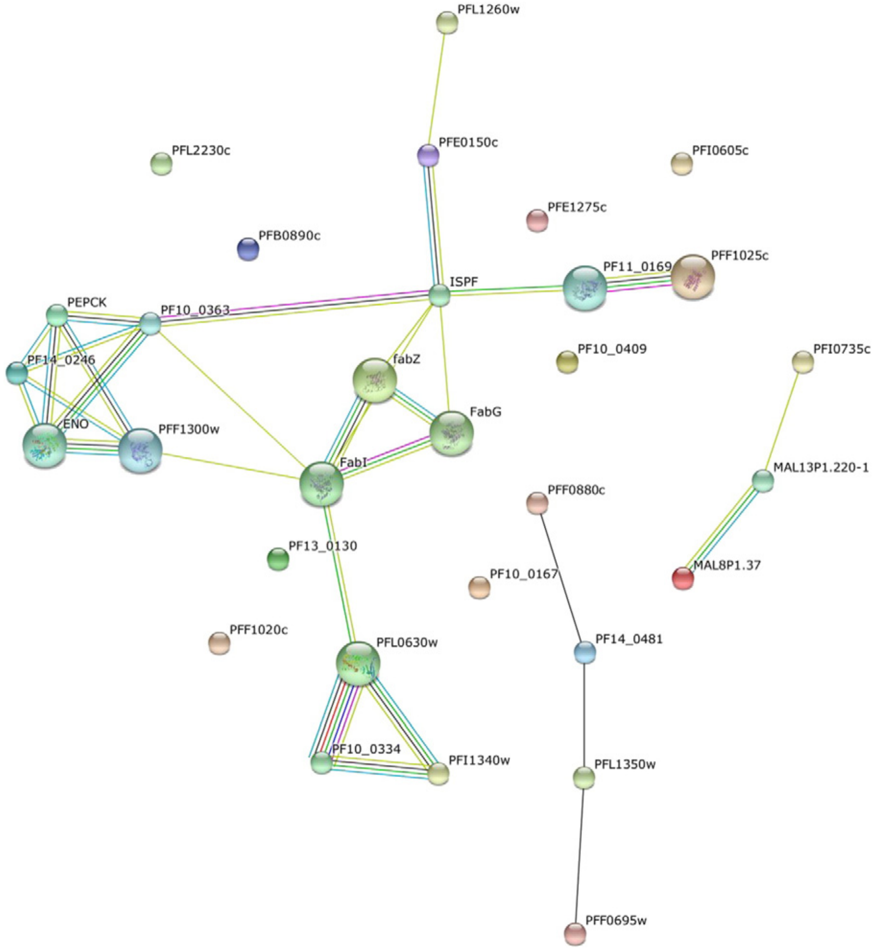

Fig. 2. Interaction network of 21 proteins *P. falciparum* by STRING database.

In the representation, network nodes are proteins, edges are the predicted functional associations. An edge association is shown in colored lines illustrating, red line — fusion evidence; a green line — neighborhood evidence; a blue line — co-occurrence evidence; a purple line — experimental evidence; a light blue line — database evidence; and a black line — coexpression evidence. A confidence score based on interactions was also calculated. Proteins with confidence score ≥ 1 are considered as highly interacting metabolic proteins. Following this interaction analysis 11 proteins are taken as highly interacting metabolic proteins (Table S3).

Drug target selection can be constricted through the prioritization of screened proteins (Table 1). In our study we focused on parameters such as molecular weight evaluation, transmembrane domain prediction *via* TMHMM server, druggability test by DrugBank, and experimental and model 3D structure determination from PDB and Mod-Base. The sub-cellular localizations of proteins were estimated through consensus results produced by ESLpred, BaCelLo and EuLoc eukaryotic sub-cellular localization prediction servers which determined 4 proteins as cytoplasmic, 2 proteins as nuclear and 5 as mitochondrial origin.

3.4. Drug targets from unique pathway

As discussed earlier the *P. falciparum* proteome contains a low degree of redundancy as a result of which most proteins participate in more than one pathway. As the proteins are filtered here through different screening processes and belong to unique pathways of pathogen they are suitably chosen as useful drug and vaccine targets against malaria parasite.

Carbon fixation in photosynthetic organism, biosynthesis of type II polyketide product, bisphenol degradation, tetracycline biosynthesis, naphthalene degradation, polycyclic aromatic hydrocarbon degradation, carbon fixation pathways in prokaryotes, aminobenzoate degradation, methane metabolism, insect hormone biosynthesis, anthocyanin biosynthesis, lipopolysaccharide biosynthesis, flavonoid biosynthesis, and carotenoid biosynthesis are some of the important unique metabolic pathways of *P. falciparum*. Fumarate hydratase, phosphoenolpyruvate carboxykinase, beta-hydroxyacyl-ACP dehydratase, phosphoenolpyruvate carboxylase, 1 putative uncharacterized protein, 1 uncharacterized protein, lipoate–protein ligase, and pseudouridine synthase are some of the non-homologous essential proteins filtered through these different computational methods that can be regarded as drug targets in malaria chemotherapy. In our analysis, we identified several proteins present in more than one pathway from the final list of 11 proteins (Table 1). These were regarded as excellent drug targets as blocking the enzyme activity will inhibit several metabolic pathways of the pathogen. As a case study, we selected “putative uncharacterized protein” and “uncharacterized protein” for further homology modeling and docking studies.

3.5. Putative uncharacterized protein and uncharacterized protein — as potential drug targets

The *P. falciparum* proteome has a vast majority of uncharacterized proteins. These need to be solved in terms of structure and function to predict their importance in pathogen metabolic activities and survival. Here, we tried to characterize two uncharacterized proteins through

Table 1
Non-homologous proteins of *P. falciparum* from PlasmoDB with reference to humans as potential drug and vaccine targets from unique pathways. The sub-cellular localization is based on the consensus results through predictions by ESLpred, BaCelLo, and EuLoc.

S. no.	Non-homologous protein targets	Associated metabolic pathways	Uniprot ID	Length (aa)	Molecular wt (in Da)	Trans-membrane domain	PDB model	Mod-Base model	Sub-cellular localization
1	Phosphoenolpyruvate carboxykinase	Carbon fixation in photosynthetic organism	Q8IDR1	583	66,209.0	0	No	Yes	Cytoplasmic
2	Pyridoxine/pyridoxal 5-phosphate biosynthesis enzyme	Biosynthesis of type II polyketide product	C6KT50	301	33,013.4	0	No	Yes	Cytoplasmic
3	Beta-hydroxyacyl-ACP dehydratase	Bisphenol degradation Biosynthesis of type II polyketide product Tetracycline biosynthesis Naphthalene degradation Limonene & pinene degradation	Q8I6T4	230	26,210.7	1	Yes	Yes	Mitochondrial
4	NADH dehydrogenase	Polycyclic aromatic hydrocarbon degradation	Q8I302	533	61,670.4	0	Yes	Yes	Mitochondrial
5	Fumarate hydratase	Carbon fixation pathways in prokaryotes	Q8I2N6	681	78,396.3	0	No	Yes	Cytoplasmic
6	Phosphoenolpyruvate carboxylase	Methane metabolism Carbon fixation in photosynthetic organism Carbon fixation pathways in prokaryotes	Q8ILJ7	1148	133,959.8	0	No	Yes	Cytoplasmic
7	Putative uncharacterized protein	Aminobenzoate degradation Stilbenoid, diarylheptanoid and gingerol biosynthesis Methane metabolism Benzoxazinoid biosynthesis Insect hormone biosynthesis Polycyclic aromatic hydrocarbon degradation Biosynthesis of 12-, 14-, 16-membered macrolides Anthocyanin biosynthesis Flavonoid biosynthesis Carotenoid biosynthesis Biosynthesis of type II polyketide product Tetracycline biosynthesis	Q8IKX0	587	68,391.3	0	Yes	Yes	Mitochondrial
8	Uncharacterized protein	Lipopolysaccharide biosynthesis	C6KT22	199	23,333.1	0	No	Yes	Nuclear
9	RNA pseudouridylate synthase	Aminobenzoate degradation	Q8I5D9	564	68,478.3	0	No	Yes	Nuclear
10	Pseudouridine synthase	Aminobenzoate degradation	O96270	338	40,297.4	1	No	Yes	Mitochondrial
11	Lipoate–protein ligase	Stilbenoid, diarylheptanoid and gingerol biosynthesis Methane metabolism Anthocyanin biosynthesis Lipopolysaccharide biosynthesis Carotenoid biosynthesis Tetracycline biosynthesis Naphthalene degradation Ethylbenzene degradation Diterpenoid biosynthesis Limonene & pinene degradation	Q8IB70	413	49,245.1	0	No	Yes	Mitochondrial

computational approaches to examine these two proteins as potential drug targets.

Putative uncharacterized protein (UniprotKB entry Q8IKX0) with a molecular weight of 68,391.3 Da is of 587 amino acids in length and provides a wide scope for drug design as it associates with aminobenzoate degradation, stilbenoid, diarylheptanoid and gingerol biosynthesis, methane metabolism, benzoxazinoid biosynthesis, insect hormone biosynthesis, polycyclic aromatic hydrocarbon degradation, biosynthesis of 12-, 14-, 16-membered macrolides, anthocyanin biosynthesis, flavonoid biosynthesis, carotenoid biosynthesis, biosynthesis of type II polyketide product and tetracycline biosynthesis of *P. falciparum*. Sequence analysis of putative uncharacterized protein showed homology to putative uncharacterized protein from *Thermus thermophilus* strain 1493, which works as a methyltransferase domain in this bacterial genome. The function of this domain is reported from the InterPro database (IPR029063). This represents a class I SAM (S-adenosyl-L-methionine)-binding-dependent methyltransferase family which is a classical methyl donor. SAM-binding methyltransferases utilize the ubiquitous methyl donor SAM as a cofactor to methylate proteins, small molecules, lipids, and nucleic acids (Sun et al., 2005).

The other uncharacterized protein (UniprotKB entry C6KT22) is of 23,333.1 Da and 199 amino acids in length can be also considered to be an essential drug target. It belongs to the lipopolysaccharide biosynthesis metabolic pathway of *P. falciparum*. This shows sequence homology to the His/Glu/Gln/Arg/opine family ABC transporter or periplasmic His/Glu/Gln/Arg/opine family-binding protein from *Silicibacter pomeroyi* presenting extracellular solute-binding protein, family III as inferred by the InterPro database (IPR001638). This protein functions in high affinity transport systems in bacterial proteomes, involved in active transport of solutes across the cytoplasmic membrane. The protein components of these traffic systems include one or two transmembrane protein components, one or two membrane-associated ATP-binding proteins (ABC transporters) and a high affinity periplasmic solute-binding protein. The latter are thought to bind the substrate in the vicinity of the inner membrane, and to transfer it to a complex of inner membrane proteins for concentration into the cytoplasm (Tam and Saier, 1993).

3.6. Homology modeling

3.6.1. Homology modeling of putative uncharacterized protein

Putative uncharacterized protein sequence of *Plasmodium* was retrieved from the Uniprot database. The query protein sequence was used for a BLASTP search against the PDB database with default parameters to select a homologous structure as template. The sequence alignment revealed structure PDB ID: 4DMG: A which is from *T. thermophilus* (Larsen et al., 2012) and has maximum identity (24%) and query coverage (75%) with query protein sequence. Sequence alignment of query and selected template was generated using Multalin (Corpet, 1988) as shown in Fig. S1. 30% sequence identity cutoff between target and template is generally considered as a threshold value for successful homology modeling (Zhexin, 2006). In our case, we considered the target-template identity value which is comparatively less than the threshold value. Therefore, we applied the alternative threading strategy by I-TASSER server to generate the 3D structure of protein. I-TASSER server is based on the secondary-structure enhanced Profile-Profile threading Alignment (PPA) and the iterative implementation of the Threading ASSEmbly Refinement (TASSER) program (Yang, 2008). However, the model generated by the homology modeling has more percentage of residues in the core allowed region (79.2%) in comparison to threading method (68.2%). So, we adopted the homology modeling technique by MODELLER software for final consideration of the structure. Among the list of five possible models, the best model was chosen based on the lowest DOPE score (−50,622.00391). DOPE or Discrete Optimized Protein Energy is a statistical potential used to assess the energy of the protein model generated through many iterations by MODELLER, which produces homology models by the satisfaction of spatial restraints. The model was further refined using the loop refinement tool provided by MODELLER v9.13. The Ramachandran plot generated by PROCHECK for the best model shows 79.2% residues in the core region and 0.2% residues in the disallowed region (Fig. S2). VERIFY3D, ERRAT and PROVE from SAVES server were also employed for evaluation of the quality of the model structure (Table S4). Fig. 3(A) shows the modeled structure of putative uncharacterized protein of *P. falciparum* with active site residues. Further, the template structure and generated model were superimposed using PyMOL visualization software with a RMS value of 0.961 Å.

3.6.2. Homology modeling of uncharacterized protein

As the “uncharacterized protein” has no structural evidence yet in the PDB structural database, a homology modeling of the protein was performed using MODELLER v9.13. Sequence similarity through a BLASTP search revealed that 3I6V: A from *S. pomeroyi* was homologous to query protein with identity (25%) and query coverage (52%) (Fig. S3). The final model was selected based on the lowest DOPE score (−13,247.29) and refined by MODELLER v9.13. PROCHECK analysis

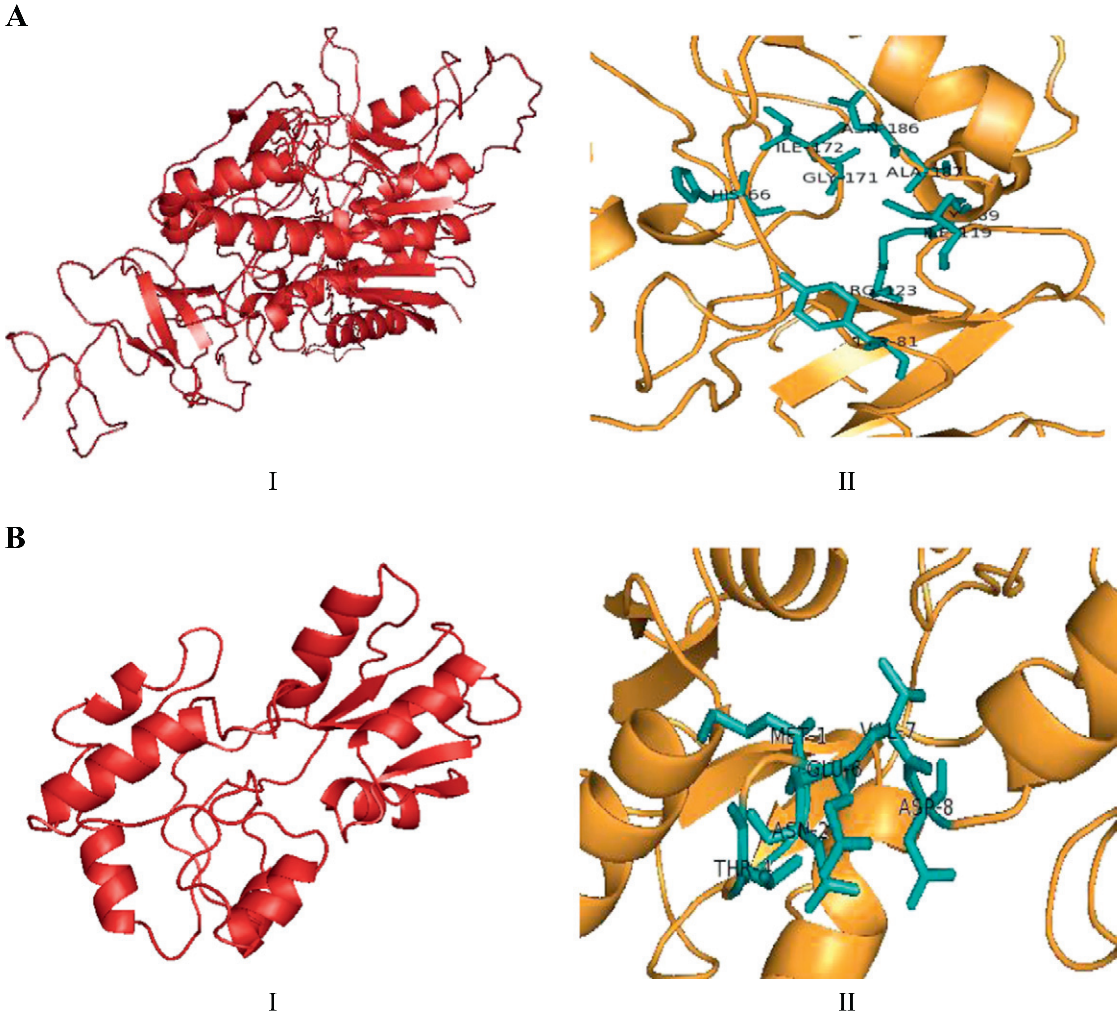

Fig. 3. Homology model: protein structures are shown in brick red color (I) and protein (bright orange color) active site residues are highlighted in cyan (II). A — Putative uncharacterized protein. B — Uncharacterized protein.

showed 87.4% residues in the most favored regions and 0.00% residues in the disallowed region of the Ramachandran plot (Fig. S4). The RMS value of the generated structure was calculated to 1.127 Å using PyMOL by structural superimposition with template. Fig. 3(B) shows the model structure with active site residue information. Through this assessment and analysis process, it is concluded that the model generated in the present study is reliable to characterize protein–substrate and protein–ligand interactions and to investigate the relation between the structure and function.

3.7. Binding site prediction and docking studies of proteins

For binding/functional site prediction here we employed the ProBiS (Protein Binding Sites Detection) database to generate hypothesis and probabilities concerning the roles of uncharacterized proteins. Potential inhibitors of target proteins for docking studies were inferred from the ZINC database through Dock Blaster. It provides a possible list of lead compounds based on their energy which can be docked to the target. The docking studies of protein were carried out using AutoDock 4 software. Receptor structure assigned with polar H-charge of the Gasteiger type and the non-polar hydrogens were merged with the carbons. The protein side chains were kept rigid in all the docking simulations. Grid maps for docking simulations were generated with 60 grid points (with 0.375 Å spacing) in the x, y, and z directions centered at the active site by the AutoGrid program of AutoDock 4. The genetic algorithm (GA) and Lamarckian genetic algorithm were assigned to the receptor and ligands by the AutoDock tool. The tool provides 10 best conformations based on their binding energy; least energy conformation was selected as best docking pose.

3.7.1. Docking of putative uncharacterized protein

In the case of putative uncharacterized protein His 66, Tyr 81, Ser 118, Ile 120, Arg 123, Gly 171, Ile 172, Asn186, Ala 187, and Gly 189 are predicted active site residues which resembles with residues Glu 102, Tyr 26, Ser 59, Val 61, Arg 64, Gly 108, Leu 109, Arg 123, Ser 124, and Gly 126 of the aligned structure from the ProBiS server. Virtual screening of putative uncharacterized protein listed ~200 potent lead compounds through the Dock Blaster software from the ZINC database (Table S5). We performed an *in silico* docking simulation study of the top five compounds from screened list using AutoDock 4 (Table S6). The docking score of the molecule (2R)-3-(5-hydroxy-1-methylbenzimidazol-2-yl)-2-(methylamino) propanoic was found to be −34.59 kcal/mol and having a H-bond interaction with Tyr 81 active site residue Fig. 4(A). This compound is proposed as a suitable inhibitor for experimental study. The next favorable interaction is displayed by the compound Ethyl with a docking score of −34.20 kcal/mol and having a hydrogen bond interaction with active site residue Tyr 81. The docking score of the third molecule, 2-(5-aminobenzimidazol-1-yl) ethanesulfonamide is reported to be −34.10 kcal/mol. The docking pose of the other screened compounds had a good docking score with hydrogen bond interaction with key active site residues. Interestingly these molecules are structurally different from each other. The result from the docking study presents an interesting opportunity for the structure-based design of a small-molecule inhibitor against the MDR strain of the *P. falciparum*.

3.7.2. Docking of uncharacterized protein

Virtual screening using the ZINC database revealed 199 lead compounds (Table S7) as inhibitors and the first five compounds were selected for docking studies. The ProBiS server helped in determining the possible binding sites of the protein through local similarity search. The target protein had a Z-score of 1.28 and alignment score of 4.79. Met 1, Asn 2, Thr 4, Glu 6, Val 7, and Asp 8 are active site residues of target protein aligned with Ile 67, Asn 68, Ala 70, Glu 72, Val 73, and Asp 74 residues of database structure after local similarity search. The receptor and ligands were docked using AutoDock 4. The docking score of the first molecule H-b-ALANYL-b-ALANINE was reported to be −68.01 kcal/mol and having a hydrogen bond interaction with Glu 6 and Val 7 residues as shown in Fig. 4(B). The second favorable interaction is displayed by H-gly-gly-ala-oh and having a hydrogen bond interaction with the active site residues Glu 6 and Val 7 and with a docking score of −66.46 kcal/mol. These molecules are potential lead compounds and the computational study will be helpful for experimental follow-up.

4. Conclusion

In the present study, a pipeline has been developed for drug discovery through a search for unique proteins by a comparative genomics study. Unique proteins are proposed to be essential drug targets. Protein–protein interaction network analysis also contributes in the selection of highly interacting metabolic proteins. We selected two uncharacterized proteins and determined the structure and the possible inhibitors through computational approaches. Results from our study could facilitate the selection of proteins and putative inhibitors for entry into drug design production pipelines in the future.

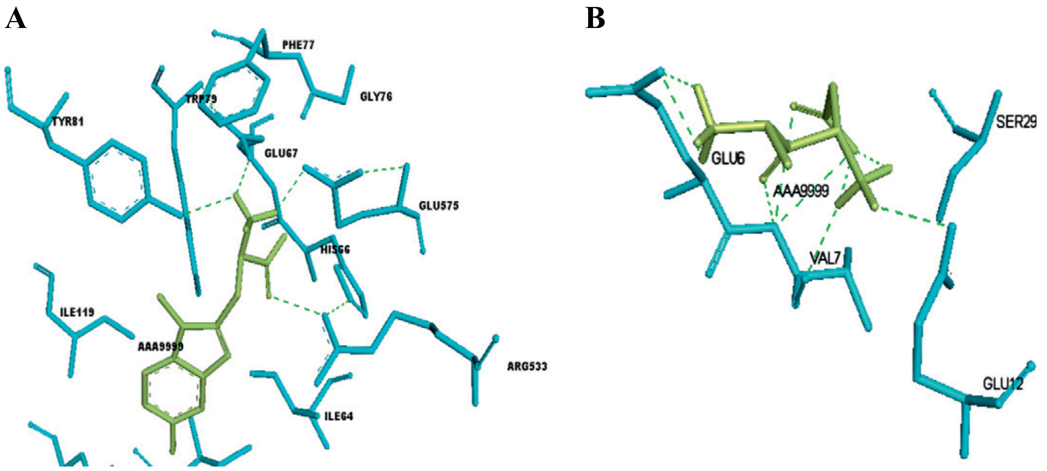

Fig. 4. Docking results: Inhibitors (green color) shows H-bond (dotted green color) with interacting residues (cyan color). A — Putative uncharacterized protein with inhibitor 76170596 (ZINC ID). B — Uncharacterized protein with inhibitor 64219287 (ZINC ID).

Acknowledgments

Miss Subhashree Rout would like to acknowledge DST, Govt. of India for the financial support to pursue her PhD work through the INSPIRE fellowship. The authors would like to acknowledge the Bioinformatics lab facility of School of Biotechnology, KIIT University during the course of the work. RKM is thankful to his wife Swagatika Mahapatra for her encouragement and support during the course of the work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mimet.2014.11.016.

References

Altschul, S.F., Thomas, L.M., Alejandro, A.S., Jinghui, Z., Zheng, Z., et al., 1997. Gapped BLAST and PSI BLAST: a new generation of protein database search programs. Nucleic Acids Res. 17, 3389–3402.

Andrews, K.T., Fisher, G., Skinner-Adams, T.S., 2014. Drug repurposing and human parasitic protozoan disease. Int. J. Parasitol.: Drugs Drugs Resist. 4, 95–111.

Anishetty, S., Pulimi, M., Pennathur, G., 2005. Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. Comput. Biol. Chem. 29, 368–378.

Arama, C., Blomberg, M.T., 2014. The path of malaria vaccine development: challenges and perspectives. J. Intern. Med. 275, 456–466.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28 (1), 235–242 (Jan 1).

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M., 1977. The Protein Data Bank. A computer-based archival for macromolecular structures. Eur. J. Biochem. 80, 319–324.

Bhasin, M., Raghava, G.P., 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res. 32 (Web Server issue), W414–W419 (Jul 1).

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, A., Estreicher, M.C., Gasteiger, E., Martin, M.J., Michoud, K., et al., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370.

Chang, T.H., Wu, L.C., Lee, T.Y., Chen, S.P., Huang, H.D., Horng, J.T., 2013. EuLoc: a webserver for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC. J. Comput. Aided Mol. Des. 27 (1), 91–103.

Chawley, P., Samal, H.B., Prava, J., Suar, M., Mahapatra, R.K., 2014. Comparative genomics study for identification of drug and vaccine targets in *Vibrio cholerae*: MurA ligase as a case study. Genomics 103, 83–93.

Chen, F., Mackey, A.J., Stoeckert Jr., C.J., Roos, D.S., 2006. OrthoMCL DB: querying a comprehensive multi-species collection of orthologous groups. Nucleic Acids Res. 34 (Database issue), D363–D368 (Jan 1).

Colovos, C., Yeates, T.O., 1993. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 2, 1511–1519.

Corpet, F., 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 16, 10881–10890.

Eisenberg, D., Lüthy, R., Bowie, J.U., 1997. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 277, 396–404.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42 (Database issue), D222–D230.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., Jensen, L.J., 2003. STRING v9.1: protein–protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41, D808–D815.

Ghosh, S., Prava, J., Samal, H.B., Suar, M., Mahapatra, R.K., 2014. Comparative genomics study for the identification of drug and vaccine targets in *Staphylococcus aureus*: MurA ligase enzyme as a proposed candidate. J. Microbiol. Methods 101, 1–8.

Irwin, J.J., Shoichet, B.K., 2005. ZINC — a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. 45 (1), 177–182.

Irwin, J.J., Shoichet, B.K., Mysinger, M.M., Huang, N., Colizzi, F., Wassam, P., Cao, Y., 2009. Automated docking screens: a feasibility study. J. Med. Chem. 52, 5712–5720.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakata, M., 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355–D360.

Konc, J., Hodoscek, M., Ogrizek, M., Konc, J.T., Janezic, D., 2013. Structure-based function prediction of uncharacterized protein using binding sites comparison. PLoS Comput. Biol. 9.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580.

Kushwaha, S.K., Shakya, M., 2010. Protein interaction network analysis — approach for potential drug target identification in *Mycobacterium tuberculosis*. J. Theor. Biol. 262, 284–294.

Larsen, L.H., Rasmussen, A., Giessing, A.M., Jogl, G., Kirpekar, F., 2012. Identification and characterization of *Thermus thermophilus* 5-methylcytidine (m5C) methyltransferase modifying 23S ribosomal RNA (rRNA) base C1942. J. Biol. Chem. 287 (33), 27593–27600.

Laskowski, R.A., MacArthur, M.N., Moss, D.S., Thorton, J.M., 1993. PROCHECK — a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283–291.

Ludin, P., Woodcroft, B., Ralph, S.A., Maser, P., 2012. *In silico* prediction of antimalarial drug target candidates. Int. J. Parasitol.: Drugs Drug Resist. 2, 191–199.

Miller, L.H., Ackerman, H.C., Su, X., Wellems, T.E., 2013. Malaria biology and disease pathogenesis: insights for new treatments. Nat. Med. 19 (2), 156–167.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J., 2009. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J. Comput. Chem. 30 (16), 2785–2791.

Mulder, N.J., Apweiler, R., 2008. The InterPro database and tools for protein domain analysis. Curr. Protoc. Bioinforma. bi0207s21 http://dx.doi.org/10.1002/0471250953 (Chapter 2: Unit 2.7).

Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., 2011. ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. 39, D465–D474.

Pierleoni, A., Martelli, P.L., Fariselli, P., Casadio, R., 2006. BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22 (14), e408–e416.

Pontius, J., Richelle, J., Wodak, S.J., 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J. Mol. Biol. 264, 121–136.

Sali, A., Blundell, T.L., 1993. Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 5, 779–815.

Sun, W., Xu, X., Pavlova, M., Edwards, A.M., Joachimiak, A., Savchenko, A., Christendat, D., 2005. The crystal structure of a novel SAM-dependent methyltransferase PH1915 from *Pyrococcus horikoshii*. Protein Sci. 14, 3121–3128.

Tam, R., Saier Jr., M.H., 1993. Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. Microbiol. Rev. 57 (2), 320–346.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B., 2003. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 31 (1), 258–261.

Wallner, B., Elofsson, A., 2003. Can correct protein models be identified? Protein Sci. 12 (5), 1073–1086.

Wiederstein, M., Sippl, M.J., 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 35, W407–W410.

World Health Organization, 2013. World Malaria Report.

Yang, Z., 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinforma. 9, 40. http://dx.doi.org/10.1186/1471-2105-9-40.

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. 14, 917–924.

Zhexin, X., 2006. Advances in homology protein structure modeling. Curr. Protein Pept. Sci. 7 (3), 217–227.