LeishBase: Leishmania major structural database

Simranjeet Kaur 1, Hitesh Patel 2, Virag Sharma 2, Prabha Garg 2, Nilanjan Roy 1,2,*

1 Dept. of Biotechnology, National Institute of Pharmaceutical Education and Research (NIPER), Punjab, India
2 Centre for Pharmacoinformatics, NIPER, Punjab, India

Submitted: 20 Jul. 2009; Accepted: 18 Sep. 2009

Abstract
Three-dimensional (3D) protein structures are of great importance for the rational design and structure-based discovery of specific inhibitors. Homology models of proteins play a significant role when no experimental three dimensional structures are available. LeishBase is a database of 347 Leishmania major proteins whose structures have been modeled by homology modeling. Detailed structural and biological details of Leishmania major proteins can provide a better understanding of their function and a way to identify potential targets involved in disease pathology. Also, an effort has been made to identify and prioritize targets among the 347 proteins present in this database. A list of top 12 targets is provided that may be probable Leishmania major drug targets. A collection of organized data under a common relational platform will give significant leverage in drug discovery effort against Leishmaniasis.

Keywords: Leishmania major, database, leishmaniasis, homology modeling.

Database URL: http://www.databases.niper.ac.in/LeishBase

INTRODUCTION

Leishmaniasis is one of the complex diseases caused by the protozoan parasites of the genus Leishmania belonging to the family Trypanosomatidae (Davies et al., 2003). Leishmania currently threatens over 12 million people in 88 countries around the world (www.who.int/zoonoses/diseases/leishmaniasis). Clinical symptoms for leishmaniasis range from the cutaneous, mucocutaneous to visceral depending on the Leishmania species (Davies et al. 2003). Cutaneous leishmaniasis is the most common form of leishmaniasis caused by species such as Leishmania major, Leishmania mexicana, Leishmania braziliensis and Leishmania panamensis. L. major genome project is completed (Ivens et al., 2005) and all the protein sequences are annotated as per existing knowledge. Present knowledge of protein structures, critical to find the small molecules having affinity to antagonize function is limited to few experimentally determined structures. A search for structures of L. major proteins in PDB could retrieve only 55 hits. Modeling the structures of annotated protein sequences, and finding new ligands using these modeled structures would be a rational way to find new drugs for Leishmaniasis.

LeishBase is a database of 347 L. major proteins modeled by Modeller (a comparative protein modeling program) using templates from Protein Data Bank (PDB). Along with structural information, significant weightage is given to general and biological information collected from various databases in LeishBase. Apart from the usual collection of information derived from the genome annotation (EC numbers, subcellular localization, GO terms), LeishBase also includes information on orthologs, paralogs, essentiality, and other pieces of information relevant for drug target validation.

Similar initiatives for creating databases for the pathogenic protozoa of trypanosomatidae family include TDR targets database (http://tdrtargets.org), SGPP (http://sgpp.org/index.shtml), TriTryDB (http://tritrypdb.org) and ProtozoaDB (http://protozoadb.biowebdb.org). TDR Targets database facilitates the prioritization of targets in complete genomes by allowing users to search for targets using defined criteria. Structural Genomics of Pathogenic Protozoa (SGPP) consortium is aimed to determine crystal structures of proteins from trypanosomatid and malaria parasites in a high throughput manner. TriTrypdb is an integrated genomic and functional genomic database for pathogens of the family...
LeishBase: Leishmania major structural database

Trypanosomatidae. ProtozoaDB hosts both genomics and post-genomics data from Plasmodium falciparum, Entamoeba histolytica, Trypanosoma brucei, T. cruzi and Leishmania major. LeishBase is developed to initially host proteins from Leishmania major only but will gradually host other trypanosomal species. LeishBase intends to be complementary to the above mentioned databases by providing further information about structural features for proteins of trypanosomatidae family. The modeled structures present in LeishBase are more accurate due to manual intervention at various stages of homology modeling as compared to TDR Targets Database in which the entries are linked to Modbase (Pieper et al., 2008) (a database of comparative protein structure models generated by an automated modeling pipeline) for getting structures. Moreover, LeishBase also provides active site information, cofactors and probable ligands for L. major proteins. In addition, Procheck Summary is provided for validation of structures. Overlapped structures of modeled proteins with corresponding templates are an added feature of this database.

For Leishmaniasis, the standard drugs like pentavalent antimonials have become obsolete due to high toxicity and emerging drug resistance while other drugs like AmphotericinB are effective but limited by their exorbitant cost. So, there is an urgent need to identify targets for development of new selective drug candidates which should have low-toxicity profile. An ideal antimicrobial enzyme target should have a fully elucidated function and mechanism of action; no human homolog, so that inhibitors are less likely to cause toxicity; and physical properties that render it amenable to high throughput biochemical assay, such as a lack of extensive hydrophobic domains and an ordered secondary structure (Brown et al., 1998). Using this approach, potential drug targets have been identified for M. tuberculosis, parasitic nematodes and arthropods (Hasan et al., 2006; Krasky et al., 2007; Kumar et al., 2007). In this work we have also identified several potential target proteins and prioritized them using in-house algorithm. These target proteins can be further taken up for experimental studies and the results could be exploited for designing antagonists active against Leishmania major.

STRUCTURE OF DATABASE AND METHODS

LeishBase contains general, biological and structural information of 347 proteins of L. major (Fig. 1). Selection criteria for proteins, various database sources,
Selection of proteins to be modeled and their corresponding templates

*L. major* proteome was downloaded from integr8 database of EBI, available at http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do (Kersey et al., 2005). Blastp (McGinnis et al., 2004) was performed against PDB database (Berman et al., 2000) on a local machine to identify template homologous proteins with available structures. Selection of the most suitable template structure and the alignment between target and template are still the predominant sources of errors in comparative models (Tramontano et al., 2003). Therefore, for template selection the blast output file was parsed such that each protein sequence should have at least 40% identity, 60% positives and 80% length ratio with the corresponding template. This exercise identified 347 proteins of *L. major* having at least one template protein in PDB fulfilling above criteria.

Homology modeling

Structures of these 347 proteins were modeled by homology modeling using MODELLER (Sali et al., 1995). The alignment of target and template sequence was performed using align2D module in MODELLER. Align2D performs global alignment of sequences using dynamic programming. Subsequently, the alignments were manually checked for errors. The optimization of the models was done using molecular dynamics (MD) with simulated annealing (SA) method in Modeller itself. These models were verified using SAVES (http://www.ncbi.nlm.nih.gov/SAVES) server (Structural Analysis and Verification Server). One of the programs in SAVES to validate the modeled structure is PROCHECK (Laskowski et al., 1993). The detailed result of SAVES is displayed under the title of “Procheck summary” in LeishBase (Fig. 1D). For each protein sequence, five models were generated initially and the model with best evaluation score was selected as final model. Manual intervention at different steps of model building can significantly improve the accuracy of the results (Paul et al., 2001). As a result, these models were found to be more accurate than those predicted by automated methods.

Identification of active site residues, ligands, cofactors and domains

Superimposition of model with template provides goodness of the homology model and also assists in positioning conserved active site residues. Active site information for template structures was obtained from Catalytic Site Atlas (CSA) database of EBI (http://www.ebi.ac.uk/thornton-srv/databases/CSA) (Porter et al., 2004). The models were overlapped with the corresponding chain of the templates to get active site information for the modeled structures. These overlapped structures were prepared using Swiss pdb viewer (Deepview) (Peitsch et al., 1997). LeishBase also provides information about small interacting molecules (ligands) of template structures from PDB (Berman et al. 2000). These ligands of template proteins could serve as a guide for putative ligand search against modeled proteins of *Leishmania*. For detailed information about domains each protein Id is linked with Pfam database (Finn et al., 2006). The cofactor information was collected from UniprotKB database (Wu et al., 2006).

General and Biological information

Apart from structural information, some essential general and biological information of these 347 proteins is included in LeishBase. General information includes physicochemical properties, classification and localization where as in biological information various pathways, homologues and post translational modifications related to these proteins are included. This information was collected from various databases like PIR (Barker et al., 2000), PROSITE (Sigrist et al., 2002), GeneDB (Hertz-Fowler et al., 2004), KEGG (Kanehisa et al., 2000) and DEG (Zhang et al., 2004). Paralogues were identified using CD-HIT program (Finn et al. 2006) with a cut off value of 60%. To identify orthologues, local blast was performed against human proteome, and best hits from human proteome were considered as orthologues. The subcellular localization of the proteins was predicted using SubCellProt (http://www.databases.niper.ac.in/SubCellProt), a tool based on machine learning approaches for assigning subcellular localizations to uncharacterized proteins (Garg et al., 2008).

IDENTIFICATION AND PRIORITIZATION OF TARGETS

Availability of a target’s structure significantly aids in rational drug design by providing strong practical advantage in high throughput docking and lead optimization studies. However to take full advantage of the database, 347 modeled structures of *Leishmania major* proteins needed to be prioritized based upon druggability. To prioritize potential drug targets a scoring scheme was implemented by taking advantage of structural and biological data present in this database. The proteins were prioritized on the basis of sequence based and knowledge based parameters which play important role in selecting a good drug target (Table 1 [Supplementary data]). The first step toward drug target identification is to identify the proteins which are unique to the pathogen and absent in the host (human) (Sakharkar 2004). This is to ensure that these enzymes
when targeted using appropriate modulators would only interfere with the pathogenic system, thereby avoiding any cytotoxic issues. This is done by comparing the pathogenic genome with the human genome and identifying proteins which do not have a human counterpart. Proteins having paralogues cannot be reliable targets because their function can be compensated for by another enzyme. Thus enzymes involved in unique essential metabolic pathways make good metabolic drug targets (Perumal et al., 2007). The scoring scheme based on these parameters identified 12 potential targets that may be considered as drug targets with more confidence.

Data collection and Methodology

All the sequence comparisons were done using the BLAST package installed on a local machine. The entire human proteome was downloaded from NCBI site. Blastp was performed against human proteome to identify those proteins that do not have any human homolog. The resulting blast output was parsed using two filters. A cut-off value of 1 e^{-10} for expectation value was used as filter I and >30% sequence identity used as filter II. To identify the essential proteins, the *Leishmania major* proteins were compared to two databases- DEG (database of essential genes) and an in-house database of fungal essential genes (http://www.databases.niper.ac.in/codaft). DEG has the essential gene data for 10 bacterial (prokaryotic) and 1 eukaryotic species. Any gene of interest which shows a good homology to the genes in the DEG is likely to be essential. The data from Database of Essential Genes (DEG) was downloaded from http://tubic.tju.edu.cn/deg/. tBlastn was performed against DEG to identify those proteins that have DEG homolog. For comparison with the eukaryotic essential genes, we have used our in-house database of fungal essential genes (CODAFT). The database contains more than 1500 essential genes and translated proteins from 4 fungal species- *Sacchromyces cerevisiae, Aspergillus niger, Candida albicans* and *Sacchromyces pombe*. KEGG database was accessed at http://www.genome.jp/kegg/genes.html to obtain the information about metabolic pathways. The protein sequences for drug targets were downloaded from Drugbank database. Blastp was performed for 347 *Leishmania major* proteins against these databases. The resulting blast files were parsed using same filters mentioned above. CD-HIT-2D program was used for identifying paralogs with 40% identity criteria for 347 proteins against *Leishmania major* proteome.

**Target scoring criterion**

To rank the output and identify most promising potential targets a scoring function was designed based on the three sequence based parameters; 1) Homology of the putative target with its human counterpart, 2) Presence of paralog of the putative target 3) Involvement of the putative target in biochemical pathway and two knowledge based parameters; 1) Essentiality of the putative target in prokaryotes and eukaryotes 2) Availability of homologue in Drug bank. Based on these five parameters the scoring function assigns a score ranging from -0.1 to 0.8 for a putative target. Each parameter is assigned some coefficient, which is its contribution to the final score for a protein. Table 1 shows the different parameters, their determination and the associated coefficients to these parameters. The parameters that are of more importance for the protein to be a good target are given a greater score. The targets were panelized with -0.2 for presence of paralog. Targets were assigned a score of 0.1 for known drug targets and essentiality while targets with no human homolog and paralog were assigned a score of 0.2. Since presence in Drugbank and essentiality are only knowledge based parameters, they are associated with lower coefficient than those associated with homology with humans and paralogy. The targets having a known metabolic pathway were assigned a score of 0.1. Using this scoring function, all of the 347 proteins were ranked and top 12 targets were obtained as best scoring targets (Fig. 2). The list of the 12 targets along with their score is shown in Table 2 [Supplementary data]. The complete list of 347 prioritized *Leishmania major* proteins is available as supplementary material. Fig. 3 shows overlapped structures of two targets with their human counterparts to make comparisons between the parasite and host protein structures.

**RESULTS**

This database is a collection of 347 protein models of *L. major* proteome along with many aspects of protein structures for the purpose of drug discovery. All the models were manually curated and verified. Comparison of host and parasite protein structures will provide insights for the identification of potential drug targets. Biologist can use it to understand interconnection between protein structure and the...
sequence. Medicinal chemist can use it to define complimentary fragments for the synthesis of inhibitor scaffold and functional groups.

Future directions

This database can be improved further by incorporating the information of proteins from other Leishmania species also. We anticipate converting it into the database for accessing any available information related to Leishmania gradually. In near future we will also incorporate programs to predict the subcellular localization, protein-protein interaction and protein folding for other Trypanosomatidae family members.

References

Albertyn J, Hohmann S, et al. (1994) GPD1, which encodes glycerol-3-phosphate dehydrogenase, is essential for growth under osmotic stress in Saccharomyces cerevisiae, and its expression is regulated by the high-osmolarity glycerol response pathway. Molecular and Cellular Biology, 14(6): 4135-4144.


LeishBase: Leishmania major structural database


