Genomics Data 7 (2016) 115–118

Contents lists available at ScienceDirect

Genomics Data

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

LmSmdB: an integrated database for metabolic and gene regulatory network in *Leishmania major* and *Schistosoma mansoni*

Priyanka Patel, Vineetha Mandlik, Shailza Singh *

National Centre for Cell Science, SP Pune University Campus, Ganeshkhind Road, Pune 411007, India

ARTICLE INFO

Article history:
Received 10 November 2015
Accepted 17 December 2015
Available online 19 December 2015

Keywords:
L.major
S.mansoni
Regulatory networks
Transcription factors
Database

ABSTRACT

A database that integrates all the information required for biological processing is essential to be stored in one platform. We have attempted to create one such integrated database that can be a one stop shop for the essential features required to fetch valuable result. LmSmdB (*L. major* and *S. mansoni* database) is an integrated database that accounts for the biological networks and regulatory pathways computationally determined by integrating the knowledge of the genome sequences of the mentioned organisms. It is the first database of its kind that has together with the network designing showed the simulation pattern of the product. This database intends to create a comprehensive canopy for the regulation of lipid metabolism reaction in the parasite by integrating the transcription factors, regulatory genes and the protein products controlled by the transcription factors and hence operating the metabolism at genetic level.

1. Introduction

An integrated relational database of curated parasites namely *L. major* and *S.mansoni,* causative organism for leishmaniasis and schistosomiasis is hosted on the NCCS webportal (http://www.nccs.res.in/LmSmdb/). This database provides information from various 'omics' studies which integrates analyze and manage environments for systems biology research. This is possible with the advent of genome analysis tool that allows identification of genes related to pathogenesis and aid to determine genes as a diagnostic marker essential in metabolic pathways which may lead to drug target identification. Towards this end, we have compared the lipid metabolic pathways specifically for *Leishmania major* and *Schistosoma mansoni* by integrating the knowledge of genome sequences of the mentioned organisms. Molecular pathways that appear to be targeted during infection, and these results highlight pathways exhibiting responses specific for a given pathogen infection. The present database is an example of the usefulness of data integration techniques which enable the creation of a hypothesis generation platform for human diseases.

Identifying a drug target and modeling a disease network is straightforward if there is conservation between two species. This conservation between the species helps to search the homologs and orthologues in a model organism. On the basis of sequence conservation, function between two genes can be deciphered which estimates probability of conservation, but evolution does not always evolve parallel with respect to evolution of sequences. We therefore made an effort to integrate two orthologues which are evolutionarily related for extracting information useful for cataloging details of interaction network particularly a gene regulatory network (GRN) and a protein interaction network (PIN) giving an insight into the relationship of protein/genes within the network. A gene regulatory network (GRN) was built in order to find the genes that are essential and highly connected with their corresponding partners. The highly connected nodes in the GRN or the hubs indicate the importance of the node in the interaction and a probable target.

In the present database, we have laid our focus on the lipid metabolism of parasites that cause infectious diseases. It is noted from [3,4] that lipid metabolism is important for maintenance of infectivity and viability of the parasite. Any change in the biochemical network results in a change in the intercellular lipid trafficking and change in the membrane composition. As lipids play an important role in the virulence and multiplication of the parasite, any approach to disrupt the lipid metabolism could result in a therapeutic strategy to stop parasite proliferation and growth [5]. In this regard, a database LmSmdB was created which studies the network of interactions, maps pathways across taxonomic branches and also incorporates data obtained from simulation studies. During the creation of LmSmdB we have simplified the process of integrating metabolic pathway interactions and protein sequence information for pathogen related studies. The main aim behind the construction of the metabolic network was to identify the important hubs which could act as drug targets in the network. We have also simplified the process of integrating metabolic pathway interactions and protein sequence information for pathogen related studies in order to address the problems related to drug target identification.

The parasite, *Leishmania major,* the causative organism for cutaneous leishmaniasis was fully sequenced in 2005 [1]. Similarly *S. mansoni*

* Corresponding author.
E-mail address: shailza_iitd@yahoo.com (S. Singh).

http://dx.doi.org/10.1016/j.gdata.2015.12.012

which causes schistosomiasis was sequenced in the year 2009. [2]. There are several databases like KEGG [8], BIOCYC [9], UNIPROT (The Uniprot Consortium) [10], NCBI [11] *etc.* that hold information about the proteins, genes and metabolites related to various biochemical pathways present in *L.major and S.mansoni.* In additions, several organism specific databases provide information on both metabolic and regulatory network whereas none of the said database provides simulation of the constructed network. In order to understand network architecture and simulation pattern of the biochemical network, we have simplified and streamlined the process of integration of molecules, genes, and protein structure in order to perform simulation of the built network. The contents on the server are exactly the same as the source code of the web page on the client side. It is the first known attempt for integrating the metabolic network and simulation of the network in a database. The different metabolic and regulatory networks were constructed using different tools [6]; the detailed information is discussed as below.

1.1. Metabolic network reconstruction

Metabolic networks of the lipid metabolism of *L.major* and *S.mansoni* genome were constructed using Cell designer 4.3 software [7], manually to curate network of the metabolism of an organism with all the genes, proteins and reactions assembled from a functionally annotated genome, biochemical data, and literature that are compiled into a stoichiometric matrix which serves as an opportunity for systematic identification of metabolic targets in the pathogen.

The regulatory network analysis of *L.major* and *S.mansoni* was dealt using the gene regulatory network interfaces with GeneNetWeaver. As the state of metabolic networks are sensed by association with the gene regulatory network, the initial data was drawn from KEGG database, Biocyc database and literature survey.

Different types of simulation like stochastic, deterministic and hybrid simulation was applied on the generated regulatory network. The advantage of using *in-silico* network is the ease to carry the perturbation experiments that can be easily simulated to produce expression data unlike *in vivo* experiments, which are usually expensive and time consuming. Moreover, both quantity and quality of the expression data generated can be controlled (e.g. by varying the amount of molecular and/or measurement noise).

2. Materials and methods

2.1. Database creation

WAMP (WAMPserver) a Windows web development environment allows the creation of web applications with Apache2, HTML and a MySQL database. WAMP 5 was used to create the LmSmdB, where HTML and PHP were used as an administrator to handle the administration of MySQL over the Web.

2.2. Querying database

The LmSmdB web page can be searched using keywords or queried using gene name or the NCBI/KEGG cluster identifier. A collection of gene aliases was assembled based on their availability on different databases. One of the key features of the database is its ability to extract data for several genes together in a batch, eliminating the need for cross referencing of the data from external databases. This batch search is useful for genomic studies where frequent updations associated with genes or proteins have to be examined frequently. The list of genes that serves as an input is in the form of a text file. This can be uploaded on the server or the list can be directly pasted into the search box. Grouping of the documents or web pages relevant to the query is done by the data sources and by the similarity searches among these documents. The entire architecture of LmSmdB database has been presented in Fig. 1. The similarity searches between two documents are based on the sequence of 'words' like '*Leishmania*', or an abbreviation, like '*Lmjr*', a sequence word 'ATGCTGGG' or 'MEPQSDVL', a number, or a combination of letters and numbers like 'Smp_124730'. A binary vector is then built on the descriptive set of words which actually define the biological

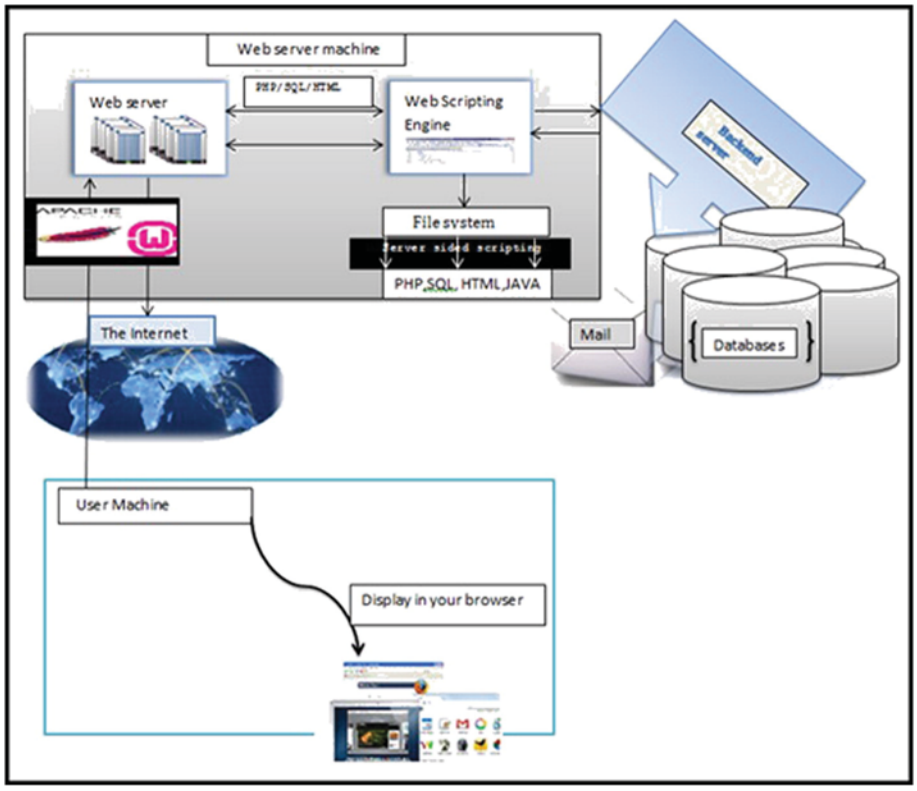

Fig. 1. The architecture of LmSmdB database consists of a backend server, web-server machine and a user machine.

information that has been searched for. Vectors with three or lesser words above threshold are removed. Certain low complexity words like 'ATGC', 'MEPQSQVL', or '0123456789', are considered as unique sets of symbols and are not considered in the vectors. The descriptive set of N words, is defined for the whole database as soon as it is updated. This set consists of 'quality' words that cover the maximum amount of documents such that N is nearly minimal with the quality of words relating to the relative stability of the corresponding component of the vector in similar documents. GIT record the changes to a file or set of files to recall a specific version.

2.3. Description and structure of database

The database architecture is provided in Fig. 1 which consists of a backend server, a web server for displaying the result, the language in which the query is submitted is in SQL and JAVA. The server sided scripting is mostly about connecting websites to back end servers which helps to enable two way communication. The first way is the server to client communication where the web pages can be assembled from back end server output and the second is the client to server communication where the user enters the information to get an output. The database consists of tables, images and links to different databases.

3. Utility and discussions

LmSmdB is a convenient, graph based data integration system that captures, incorporates, and manages available data related to the functional importance of several genes and proteins. The database is structured in the form of a network that incorporates features of a given entity such as genes, proteins and pathways. The database is also capable of dynamically incorporating new sets of objects and their relation thus integrating data types like tables and sequences.

The protein information along with the compartmentalization is included in the database which links to the KEGG database. The GRN and the PIN are provided in the image format, the protein and genes included in the table format along with the link to other databases (KEGG). The outline of the database along with the work-flow is shown in Fig. 2.

The database architecture consists of a back end server and a web server for displaying the result. The language in which the query is submitted is in SQL, HTML and PHP. Server side script connects the websites

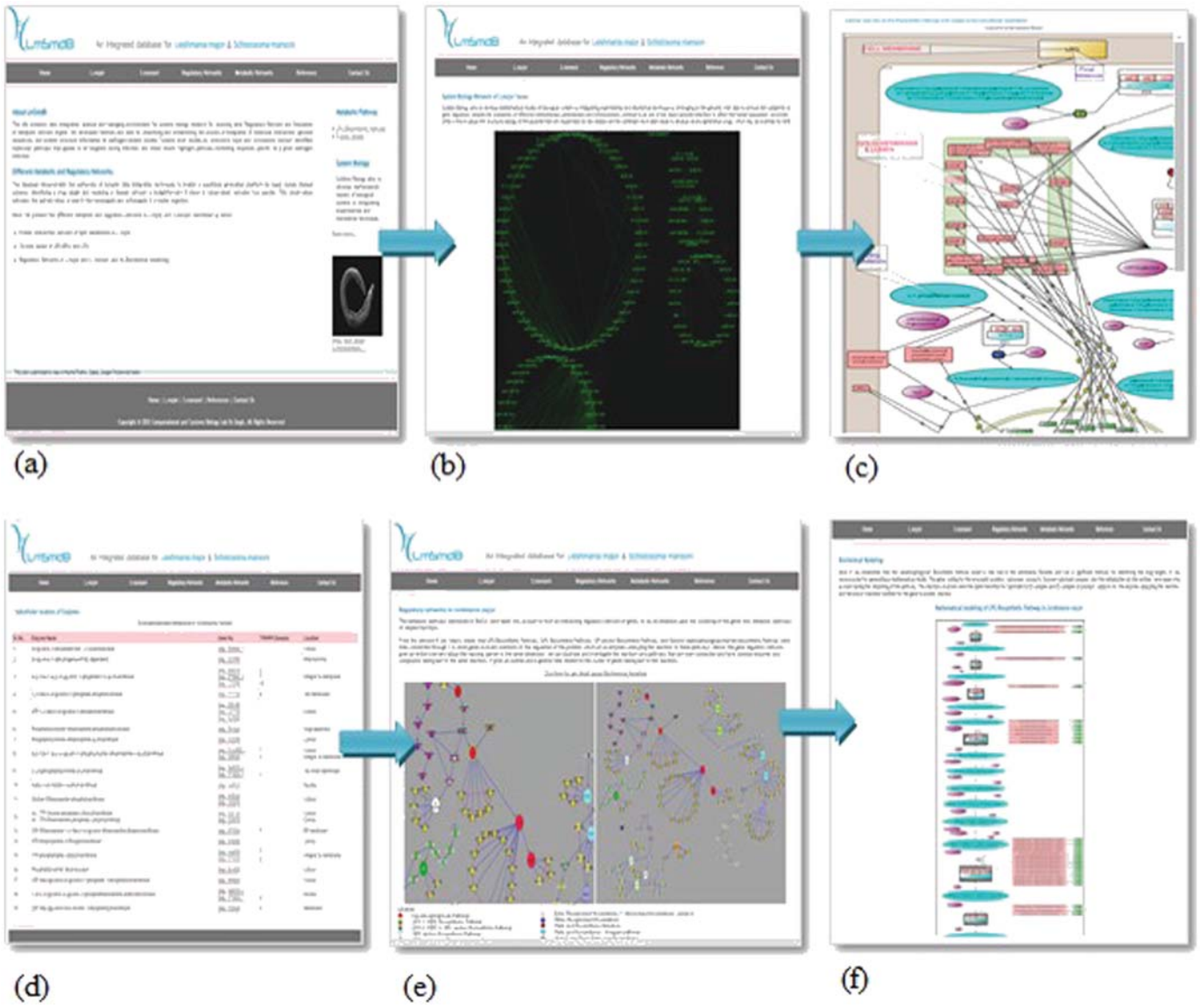

Fig. 2. LmSmdB database report page (a) home page of the database, gives information on the disease leishmaniasis and schistosomiasis, metabolic networks, regulatory networks, references and contact details. (b) GRN of *Schistosoma mansoni*. (c) Metabolic network reconstruction for *L. major*. (d) List of proteins in the network. (e) The Regulatory network of *L. major*. (f) The Metabolic network reconstruction in *Schistosoma mansoni*.

to back end server to enable two way communication *i.e.* from client side where the web pages are assembled and the second is the client to server communication where the user enters the information to get an output. The internal database of the system is structured in the form of a network database which contains all the features assigned to bio entities like genes, proteins and pathways. New sets of objects along with their relations within a search are easily incorporated. This is done by integrating several data types like graphs, tables and sequences. Any problem arising in the data integration are dealt with *via* an ontology driven data mapping, multiple data annotation and heterogeneous data querying in order to enable integration of the user's data. The graphs obtained in Cell designer, the interactions are visualized in JAVA which depicts how a gene, metabolite and a protein are connected. We simplified and effectively integrated the molecular interaction in the form of protein as nodes and edges as interaction together with the structural information. Systems level studies on *Leishmania major* and *Schistosoma mansoni* identified molecular pathways that appear to be targeted during infection, and these results highlight pathways exhibiting responses specific for a given pathogen infection demonstrating the practical use of data integration techniques, to enable hypothesis generation platform for major human disease systems providing access to heterogeneous information that are of value to researchers targeting the tropical diseases.

4. Availability

LmSmdB is hosted on the webportal of National Centre for Cell Science (NCCS).

5. Data maintenance

LmSmdB will be continuously updated, through manual screening of new publications and literature review. In addition, we also openly welcome suggestions by researchers that include new or missing findings to be inserted in the database by contacting authors by email. Our contact information is reported in the 'Contact Us' page.

6. Conclusion

LmSmdB is the first online resource containing genomic information related to metabolic and regulatory network specifically for *L. major* and *S. monsoni*. It is designed as a common purpose framework for systems biology providing formalized graphic notation of biological systems structure and functioning, their visualization and simulations as well as access to databases with relevant data. A user can browse information regarding the disease caused by these two pathogens and the gene involved in the pathogenesis. By looking at LmSmdB, the user can also get the link to the KEGG pathway of particular gene, which is directly involved in the pathway or network.

Acknowledgments

The authors would like to thank Department of Biotechnology, Ministry of Science and Technology, Government of India for funding the work. (BT/PR3140/BID/7/379/2011). The authors would also like to thank the Director, National Centre for Cell Science (NCCS), Pune for supporting the Bioinformatics and High Performance Computing Facility at NCCS, Pune, India.

References

[1] A.C. Ivens, et al., The genome of the kinetoplastid parasite, *Leishmania major*. Science (5733) (2005) 436–442.
[2] M. Berriman, et al., The genome of the blood fluke *Schistosoma mansoni*. Nature 460 (7253) (2009) 352–358.
[3] Niemela, et al., Bioinformatics and computational method for lipidomics. J. Chromatogr. B 877 (2009) 2855–2862.
[4] Chavali, et al., Systems analysis of metabolism in the pathogenic trypanosomatid *Leishmania major*. Mol. Syst. Biol. 4 (2008) 1–19.
[5] S. Singh, S. Shinde, in: D. Ekinci (Ed.), Stochastic simulation for biochemical reaction networks in infectious disease, medicinal chemistry and drug design, InTech, ISBN: 978-953-51-0513-8, 2012.
[6] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11) (2003) 2498–2504 (Nov).
[7] A. Funahashi, N. Tanimura, M. Morohashi, H. Kitano, CellDesigner: a process diagram editor for gene-regulatory and biochemical network. Biosilico 1 (2003) 159–162.
[8] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40 (2012) D109–D114.
[9] R. Caspi, T. Altman, J.M. Dale, K. Dreher, C.A. Fulcher, F. Gilham, P. Kaipa, A.S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L.A. Mueller, S. Paley, L. Popescu, A. Pujar, A.G. Shearer, P. Zhang, P.D. Karp, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome. Nucleic Acids Res. 38 (2010) D473–D479 (2010).
[10] The Uniprot consortium, Reorganizing the protein space at the Universal Protein Resource Uniprot. Nucleic Acids Res. 40 (2012) D71–D75.
[11] L.Y. Geer, A. Marchler-Bauer, R.C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, S.H. Bryant, The NCBI BioSystems database. Nucleic Acids Res. D492-6 (2010).