A comparative protein function analysis database of different *Leishmania* strains

Manas Ranjan Dikhit, Yangya Prasad Nathasharma, Lelin Patel, Sindhu Prava Rana, Ganesh Chandra Sahoo*, Pradeep Das

BioMedical Informatics Division, Rajendra Memorial Research Institute of Medical Sciences (RMRIMS), Agam kuan, Patna-800007, India; Ganesh Chandra Sahoo - Email: ganeshiitkgp@gmail.com; Phone: 0612 -2631565; Fax: 0612 -2634379; *Corresponding author

Received July 26, 2010; Accepted August 26, 2010; Published March 02, 2011

Abstract:
A complete understanding of different protein functional families and template information opens new avenues for novel drug development. Protein identification and analysis software performs a central role in the investigation of proteins and leads to the development of refined database for description of proteins of different *Leishmania* strains. There are certain databases for different strains that lack template information and functional family annotation. Rajendra Memorial Research Institute of Medical Sciences (RMRIMS) has developed a web-based unique database to provide information about functional families of different proteins and its template information in different *Leishmania* species. Based on the template information users can model the tertiary structure of protein. The database facilitates significant relationship between template information and possible protein functional families assigned to different proteins by SVMProt. This database is designed to provide comprehensive descriptions of certain important proteins found in four different species of *Leishmania* i.e. *L. donovani*, *L. infantum*, *L. major* and *L. braziliensis*. A specific characterization information table provides information related to species and specific functional families. This database aims to be a resource for scientists working on proteomics. The database is freely available at http://biomedinformri.org/calp/.

Keywords: *Leishmania*, database, protein function, functional family, CPL, Computational proteomics of *Leishmania*

Background:
Kala-azar or Leishmaniasis is identified by clinical syndromes caused by obligate intracellular protozoa of the genus *Leishmania* and transmitted from one host to another by the bite of blood sucking sand fly vectors [1]. The genomes of three species have been sequenced. There are relatively few species-specific differences in gene content between the sequenced genomes, but nearly 8% of the genes appear to be evolving at different rates [2]. Knowledge about protein function is essential in the understanding of biological processes [3]. No computational functional analysis of different proteins of *Leishmania* is available till date. As the gap between the amount of sequence information and functional characterization widens, increasing efforts are being intended for the construction of databases. For scientist, it is therefore helpful to have a single data collection point, which integrates research interrelated data from diverse domains. Large scale of protein sequences is available at the National Center for Biotechnology Information (NCBI) protein database [4] and supplementary data in the published literature. In silico analysis gives us an idea on the role of different proteins in replication, survival and spread in the host [5]. Computational proteomics of *Leishmania* (CPL) involves the general tasks related to analysis of any novel sequences, such as functional annotation and template information of the sequences. Support vector machine (SVM) is a useful classifier for predicting the functional classes of distantly related proteins [6, 7]. The function of a protein depends on its tertiary structure. The structure and function of a protein gives much more insight of the protein than its sequence [8]. Structural genomics are yielding many protein structures that have unknown function. Nevertheless, successive experimental investigation is costly and time-consuming, which makes computational methods for predicting protein function very attractive [9].

Therefore, a number of methods for the computational prediction of protein structure from its sequence have been proposed. The simulated model of the protein structure refers to the construction of an atomic-resolution model of the target protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein (i.e. template) [10]. The critical first step in homology modeling is the discovery of the best template structure based on which a tertiary structure will be modeled [11]. Considering the biological significance of protein and with the aim of providing easy access to large and growing volume of data, we have developed a repository for most common proteins in which user can get the information about the template and functional family of protein. As drug resistance problem persists in case of Leishmaniasis, template information will help further modeling and analysis of different essential proteins which would lead to the discovery of novel lead compounds.

Methodology:
The large scale of protein sequences have been reported in the NCBI protein database and supplementary data in the published literature. The commonly available virulent sequences of *Leishmania* have been downloaded from the National Center for Biotechnology Information (NCBI) and GeneDB [12]. Different strains of the same species from samples collected from diverse location or at different times may have completely identical sequences. Redundancy and repetition in protein sequences has been carefully removed by using ALIGN software to obtain a unique dataset [13]. Exactly matching sequences taken from multiple sources were eliminated while constructing the dataset. The raw dataset was preprocessed to remove the sequence smaller than 50bp while analyzing with different software.
Database design:
A rational database was constructed in MySQL for storage and query of data. It includes two key entities namely molecular function and template structure which fetches the probable function and most appropriate virtual structure of the protein. The database consists of three layers: the basal layers’, ‘Application layer’ and ‘UI’ layer. These layers are developed using Php, CSS and JavaScript. The information’s are managed in protein level to provide timely and general view of the data. The data and information have been stored in MySQL relational database. Meta information for different types of biological data is placed as individual table in this layer (Figure 1).

Results and Discussion:
Identification of diverse protein functions may facilitate a mechanistic consideration. Our study from SVMProt suggests that the proteins of different functional family to find out the proteins that possesses the same function (Figure 2). The user can compare the proteins of different species and their functional families. The user can also compare tertiary structure of the templates (Figure 3).

Utility:
With the aim of providing easy access to large and growing volume of data, a database of most common protein is developed. This is the first web resource which provides the common protein sequence of four strains as well as their functional classes for comparison. The database has been analyzed, organized and integrated to develop a user friendly interface. The web interface enables the user to execute a quick and efficient search and comparison. The database will be an extremely useful resource for computational and experimental biologists working in Leishmania proteomics and related areas.

Acknowledgements:
We thank Dr. Meera Singh of ICMR, New Delhi for helping us in setting up our new biomedical informatics department in RMRIMS, Patna, India. The authors would like to thank Mr. Rajani Kanta Nahak, Mr. Kanhu Charan Maharan, Miss Smita Sahoo, Miss Chanda Jha and Miss Mukta Rani for helpful discussion and valuable suggestions.

References:
[1] Sahoo GC et al. J Proteomics Bioinform. 2009 2(1): 32
[2] Peacock CS et al. Nat Genet. 2007 39: 839 [PMID: 17572675]
[3] Dikhit MR et al. International Journal of Biometrics and Bioinformatics 2009 3(4): 59
[4] http://www.ncbi.nlm.nih.gov
[5] Dikhit MR et al. Bioinformatics 2009 3(7): 299 [PMID: 19293996]
[6] http://jmg.ez3.nus.edu.sg/cgi-bin/svmprot.cgi
[7] Sahoo GC et al. Bioinformation 2008 3(1): 1 [PMID: 19052658]
[8] Laskowski RA et al. Nucleic Acid res. 2005 33: W9/W9 [PMID: 15908588]
[9] Lee D et al. Nat Rev Mol Cell Biol. 2008 8: 995 [PMID: 18037900]
[10] Bordoli L et al. Nat Protoc. 2009 4(1): 1 [PMID: 19131951]
[11] Ozlem TB et al. South African Journal of Science 2008 104(1): 2
[12] http://www.genedb.org/Homepage
[13] http://www.ebi.ac.uk/Tools/emboss/align/index.html

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Citation: Dikhit et al. Bioinformation 6(1): 20-22 (2011)
### Supplementary material:

| Functional family                        | *L. braziliensis* | *L. infantum* | *L. donovani* | *L. major* |
|------------------------------------------|------------------|---------------|---------------|------------|
| Zinc-binding                             | ✓                | ✓             | ✓             | ✓          |
| DNA repair                               | ✓                | ✓             | ✓             | ✓          |
| RNA-binding Proteins                     | ✓                | ✓             | ✓             | ✓          |
| Metal-binding                            | ✓                | ✓             | ✓             | ✓          |
| Lyases - Carbon-Carbon Lyases            | ✓                | ✓             | ✓             | ✓          |
| All DNA-binding                          | ✓                | ✓             | ✓             | ✓          |
| Outer membrane                           | ✓                | ✓             | ✓             | ✓          |
| Aptamer-binding protein                  | ✓                | ✓             | ✓             | ✓          |
| All lipid-binding proteins               | ✓                | ✓             | ✓             | ✓          |
| Actin binding                            | ✓                | ✓             | ✓             | ✓          |
| Lyases - Carbon-Oxygen Lyases, Actin binding | ✓            | ✓             | ✓             | ✓          |
| Oxidoreductases - Acting on the CH-CH group of donors | ✓      | ✓             | ✓             | ✓          |
| Manganese-binding                        | ✓                | ✓             | ✓             | ✓          |
| DNA recombination                        | ✓                | ✓             | ✓             | ✓          |
| mRNA slicing                             | ✓                | ✓             | ✓             | ✓          |