BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 8(24)

MycoProtease-DB: Useful resource for *Mycobacterium tuberculosis* complex and nontuberculous mycobacterial proteases

Lingaraja Jena, Satish Kumar & Bhaskar Chinnaiah Harinath*

Bioinformatics Centre, JB Tropical Disease Research Centre, Mahatma Gandhi Institute of Medical Sciences, Sevagram (Wardha) 442102, Maharashtra, India; Bhaskar Chinnaiah Harinath – Email: bc_harinath@yahoo.com; Phone: +91 7152 – 284341- 284355, Ext: 262, 303; Tele Fax: (07152) 284038; *Corresponding author

Received November 16, 2012; Accepted November 18, 2012; Published December 08, 2012

Abstract:
MycoProtease-DB is an online MS SQL and CGI-PERL driven relational database that domiciles protease information of *Mycobacterium tuberculosis* (MTB) complex and Nontuberculous Mycobacteria (NTM), whose complete genome sequence is available. Our effort is to provide comprehensive information on proteases of 5 strains of *Mycobacterium tuberculosis* (H_{37}Rv, H_{37}Ra, CDC1551, F11 and KZN 1435), 3 strains of *Mycobacterium bovis* (AF2122/97, BCG Pasteur 1173P2 and BCG Tokyo 172) and 4 strains of NTM (*Mycobacterium avium* 104, *Mycobacterium smegmatis* MC2 155, *Mycobacterium avium paratuberculosis* K-10 and *Nocardia farcinica* IFM 10152) at gene, protein and structural level. MycoProtease-DB currently hosts 1324 proteases, which include 906 proteases from MTB complex with 237distinct proteases & 418 from NTM with 404 distinct proteases. Flexible database design and easy expandability & retrieval of information are the main features of MycoProtease-DB. All the data were validated with various online resources and published literatures for reliable serving as comprehensive resources of various Mycobacterial proteases.

Availability: The Database is publicly available at http://www.bicjbtdrc-mgims.in/MycoProtease-DB/

Keywords: Mycobacterium tuberculosis complex, Database, Protease, NTM

Background:
Tuberculosis continues to be a major health problem worldwide and it is estimated that in 2011, nearly 8.7 million new cases of TB with 1.4 million deaths among HIV-negative people and an additional 0.43 million deaths from HIV-associated TB **[1]**. It has been found that proteases of *Mycobacterium tuberculosis* have an important role in pathogenesis of the organism **[2]**. Mehaffy *et al* (2012) in their review on MTB proteomics, highlighted the role of proteases in the virulence and pathogenicity of human pathogens **[3]**. Nontuberculous Mycobacteria (NTM) can also produce localized disease in the lungs, lymph glands, skin, wounds or bone **[4]**. So, our effort is to explore proteases of MTB complex and NTM at gene, protein and structural level.

Methodology:
Database Architecture & Design
The relational database was developed using Microsoft SQL Server 2005 as the back end. The website is powered by Apache HTTP Server 2.2.6. HTML, JavaScript and CGI-PERL based web interfaces have been developed which dynamically execute the SQL queries. The MycoProtease-DB data and related information are stored in MS SQL relational database tables.

Data Curation
Twelve Mycobacterial strains (Eight MTB complex and four NTM) were identified whose complete genome sequences were available at National Centre for Biotechnology Information

 © 2012 Biomedical Informatics

(NCBI) **[5]** as follows: 5 strains of *Mycobacterium tuberculosis* ($H_{37}Rv$, $H_{37}Ra$, CDC1551, F11 and KZN 1435), 3 strains of *Mycobacterium bovis* (AF2122/97, BCG Pasteur 1173P2 and BCG Tokyo 172) and 4 strains of NTM (*Mycobacterium avium* 104, *Mycobacterium smegmatis* MC2 155, *Mycobacterium avium paratuberculosis* K-10 and *Nocardia farcinica* IFM 10152). Protease information was collected from MEROPS **[6]** followed by NCBI, UniProt **[7]**, Kyoto Encyclopedia of Genes and Genomes (KEGG) **[8]**, TubercuList **[9]** and published literatures for individual strain. Protease length, molecular weight, theoretical isoelectric point was calculated using ExPASy ProtParam **[10]** tool. Then, all the curated information of each protease was inserted into MycoProtease-DB. Standalone BLAST **[11]** was used for obtaining homologous sequences of each protease and corresponding homologous IDs were also added in the database table.

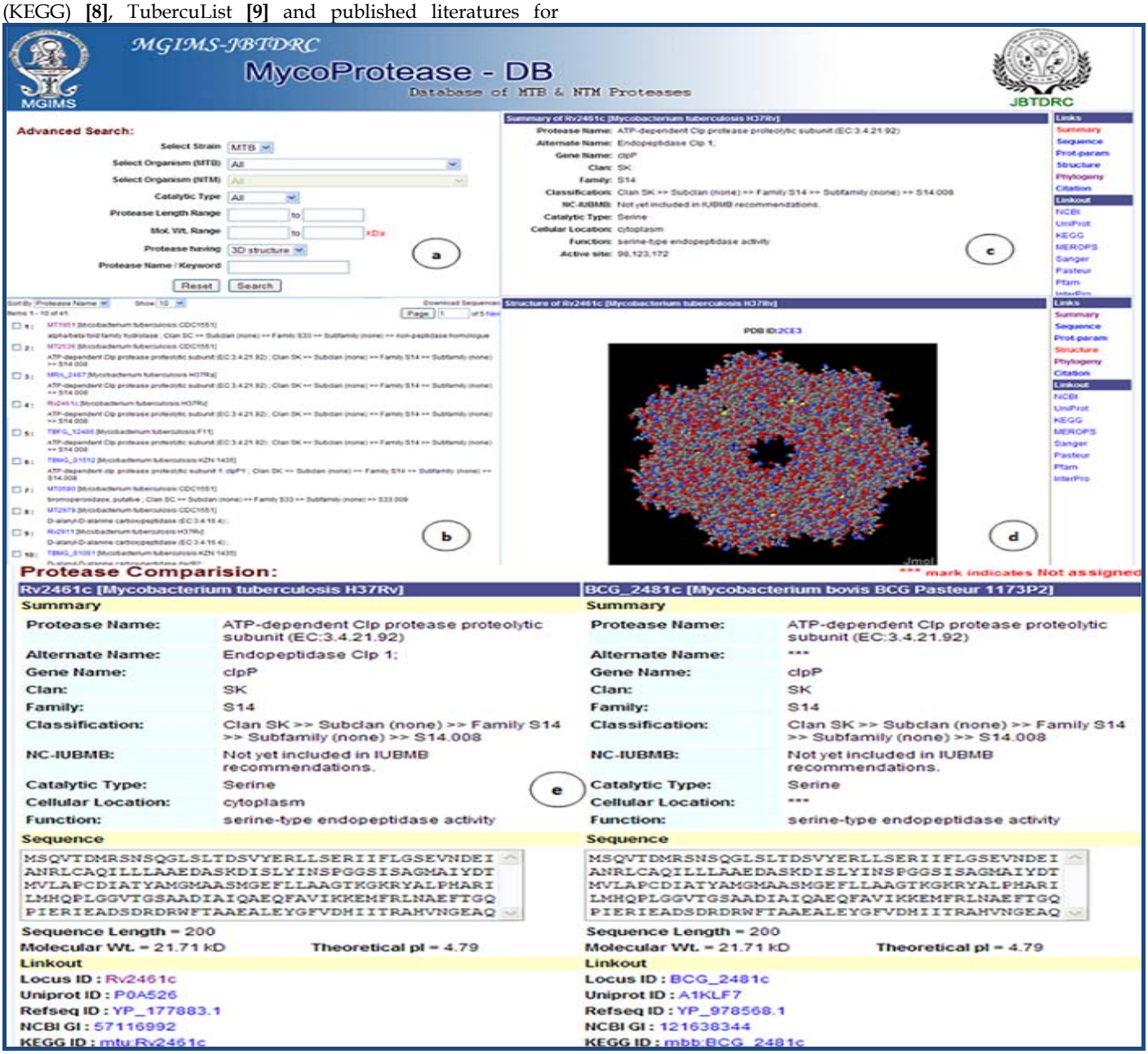

Figure 1: MycoProtease-DB snapshots **(a)** Advanced search Page; **(b)** Search result; **(c)** Summary page; **(d)** Structure visualization page powered by Jmol; **(e)** Protease Comparison page.

Data Access

The interfaces in MycoProtease - DB are designed in a manner to help users in easy navigation and retrieve information from database (Figure 1). The database can be queried to obtain the protease information in many ways through a user friendly web interface as follows. i) The user can enter the desired protease name to access the Meta information about proteases. The user can also search by catalytic type, amino acid length, molecular weight, NCBI GI, RefSeq, UniProt, KEGG, Locus ID etc. ii) Advanced search option is provided for searching more user specific information regarding proteases. Using this option, user can search protease information according different strains, catalytic type, specific protease length & molecular weight range etc. There is also an option for downloading selected sequences in fasta format. iii) A dynamic result page appears after any search in which user can sort the searched

© 2012 Biomedical Informatics

result (protease list) by name, catalytic type, molecular weight and sequence length. The user can also restrict the no of items to be shown per page obtained in searched result. iv) Along with Summary information (Name, Gene, Clan, Family, Catalytic type, Cellular location, Function etc.) each protease entry has also Sequence information (amino acid sequences, length, molecular weight, theoretical isoelectric point (pI), nucleotide sequence & length and related homologous ids), Protease parameters (Amino acid length, composition, molecular wt, pI, atomic composition, formulae etc), Phylogeny (Multiple Sequence Alignment [MSA] & Phylogenetic tree of homologous sequences) by GeneBee - Molecular Biology Server **[12]**, Structure (3D structure if available viewed by Jmol **[13]**, Citation and Linkout (Links to NCBI, UniProt, KEGG, MEROPS, Pfam, InterPro of corresponding protease entry). v) Protease Comparison: Provides protease - protease comparison in "Tools & Analysis" page. The user can enter the corresponding ID either of NCBI-GI, locus, RefSeq, Uniprot, KEGG ID to compare two proteins. Tools for calculating protein parameters, Residues finder are also integrated in "Tools & Analysis" page.

Comparison with other Databases

Presently, MEROPS, the database of peptidases, contains protease information of 8546 organisms. In addition, protease data is available at NCBI, UniProt, KEGG and Tuberculist databases but they are not specific and comprise of huge data of other organisms also. MycoProtease-DB is a comprehensive database with information on Mycobacterial proteases.

Utility:

MycoProtease-DB is a comprehensive database on proteinases of 8 MTB complex and 4 NTM strains. It has total of 1324 (641 distinct) peptidases, which include 906 proteases from MTB complex with 237 distinct & 418 from NTM with 404 distinct proteases. This information facilitates further analysis of MTB and NTM proteases in molecular and functional level. It will be supportive to the researchers to carry out further work in this field.

Caveats:

MycoProtease-DB does not include protease information of all mycobacterial strains as they are not completely sequenced. There are 154 hypothetical proteins with protease activity in MycoProtease-DB which are yet to be annotated.

Future Developments:

As and when in future, new mycobacterial strains are sequenced and protease data are available in public databases; we shall continue to update MycoProtease-DB including annotated information of hypothetical proteases.

Acknowledgement:

This study was supported by Department of Biotechnology, Ministry of Science & Technology, Govt. of India. Authors convey thanks to Shri Dhiru S Mehta, President, KHS for his keen interest and encouragement.

References:

[1] http://www.who.int/tb/publications/global_report/en/
[2] Upadhye V *et al*. *Scand J Infect Dis.* 2009 **41**: 569 [PMID: 19479636]
[3] Mehaffy MC *et al*. *J Proteome Res*. 2012 **11**: 17 [PMID: 21988637]
[4] Katoch VM, *Indian J Med Res*. 2004 **120:** 290 [PMID: 15520483]
[5] http://www.ncbi.nlm.nih.gov/genome/
[6] Rawlings ND *et al*. *Nucleic Acids Res.* 2012 **40:** D343 [PMID: 22086950]
[7] UniProt Consortium, *Nucleic Acids Res.* 2010 **38**: D142 [PMID: 19843607]
[8] Kanehisa M & Goto S, *Nucleic Acids Res.* 2000 **28**: 27 [PMID: 10592173]
[9] http://tuberculist.epfl.ch/
[10] http://web.expasy.org/protparam/
[11] Altschul SF *et al*. *J Mol Biol.* 1990 **215**: 403 [PMID: 2231712]
[12] http://www.genebee.msu.su/services/malign_reduced.html
[13] http://jmol.sourceforge.net/

Edited by P Kangueane

Citation: Jena *et al.* Bioinformation 8(24): 1240-1242 (2012)

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