Tuberculosis 91 (2011) 556–562

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

MOLECULAR ASPECTS

MtbSD—A comprehensive structural database for *Mycobacterium tuberculosis*

Sameer Hassan [a], P. Logambiga [a], A. Mohan Raman [a], T.K. Subazini [b], V. Kumaraswami [a], Luke Elizabeth Hanna [a,*]

[a] National Institute for Research in Tuberculosis, Chetpet, Chennai 600 031, India
[b] AU-KBC Research Centre, Madras Institute of Technology, Chennai 600 044, India

ARTICLE INFO

Article history:
Received 26 April 2011
Received in revised form
6 August 2011
Accepted 8 August 2011

Keywords:
Mycobacterium tuberculosis
Proteins
Ligands
Domains
Database

SUMMARY

The *Mycobacterium tuberculosis* Structural Database (MtbSD) (http://bmi.icmr.org.in/mtbsd/MtbSD.php) is a relational database for the study of protein structures of *M. tuberculosis*. It currently holds information on description, reaction catalyzed and domains involved, active sites, structural homologues and similarities between bound and cognate ligands, for all the 857 protein structures that are available for *M. tb* proteins. The database will be a valuable resource for TB researchers to select the appropriate protein–ligand complex of a given protein for molecular modelling, docking, virtual screening and structure-based drug designing.

1. Introduction

Mycobacterium tuberculosis, the etiological agent for tuberculosis, causes approximately 8–10 million new infections and 3 million deaths worldwide every year.[1] In 1993, World Health Organisation (WHO) declared tuberculosis a global emergency. TB is one of the leading causes of mortality in India, killing 2 persons every 3 min, nearly 1000 every day.[2]

The complete genome of *M. tuberculosis* comprising of 4,411,529 bp and around 4000 genes was sequenced in 1998.[3] Availability of the mycobacterial genome sequence and advancement in structural genomics have set up a platform to answer questions such as the functioning of the organism as an integrated system and its activity in conjunction with the host. Three-dimensional structures of proteins are important for understanding their biological function as well as their interaction with ligands. Structure determination for hypothetical proteins could help in the identification of the biological function of a particular protein based on clues obtained from proteins with even distant structural homology and no apparent sequence identity.[4,5]

Protein Data Bank (PDB) is a valuable resource of structural data for proteins of all eukaryotes and prokaryotes. TBSGC, also hosts similar information but is specific for *M. tuberculosis*. Currently, both databases contain 853 protein structures for 328 gene products of *M. tuberculosis*, resulting from multiple solved structures for many of these proteins, viz. mutant forms, structures for individual or multiple domains, and complexes with different ligands. Since proteins are flexible molecules capable of changing their structure based on external stimuli, binding of ligands may induce certain changes in the structure.[6] These changes may be important for protein function. Exploring the changes that have occurred in the protein structure due to ligand binding will further help in understanding the functions of the proteins, since some of these structural changes may be important for protein function.

When ligand-bound protein conformations are not available, structure-based drug design becomes highly challenging. Several studies have shown that virtual screening with an apo structure usually results in a poor enrichment factor compared to screening with holo structure even when the structural difference between the two is small.[7–9] X-ray or NMR structures of a protein/enzyme are not always in complex with its natural substrate or product. Binding of ligands other than cognate ligands could also bring about a range of structural deviations in the proteins. These changes can result in poor enrichment factor. Identification of the appropriate protein–ligand complex based on structural similarity between the bound and cognate ligands is very essential for proper modelling, docking, as well for drug designing studies.

Further, in order to understand the catalytic activity of a target protein, availability of its crystal structure in combination with its

* Corresponding author. Scientist C, Dept. of Clinical Research, National Institute for Research in Tuberculosis, Chetpet, Chennai 600031, India. Tel.: +91 44 28369575; fax: +91 44 28362528.
E-mail address: hanna@trcchennai.in (L.E. Hanna).

1472-9792/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tube.2011.08.003

appropriate ligand is very essential. At present, information on catalytic activity and similarity between cognate and bound ligands are available in SwissProt[10] and PROCOGNATE[11] databases respectively. Though Swissprot and other databases such as TBSGC,[12] Tuberculist[13] and TBDB[14] contain information on *M. tuberculosis* protein structures, none of these databases provide systematic grouping of structures for each protein, or highlight the differences between structures in the context of domain coverage, bound ligands etc.

In this scenario, users have to visit multiple databases and employ many tools to acquire complete information on a given protein for protein modelling, docking and structure-based drug-designing studies. Keeping in mind the need for a complete structural database for *M. tuberculosis* proteins, that would incorporate all the above mentioned features, we developed a database called MtbSD (http://bmi.icmr.org.in/mtbsd/MtbSD.php).

2. Materials and methods

2.1. MtbSD platform

MySQL server version 5.1.41 was used in MtbSD to store, retrieve and manage data. All the scripts for data querying and retrieving were written in PHP, a highly efficient and widely used scripting language. The web interfaces were designed using HTML language and Cascading Style Sheet (CSS) for consistent styling.

2.2. Database generation

Sequence of all *M. tuberculosis* proteins was retrieved from NCBI database and searched against PDB[15] using the BLAST[16] tool. Information about different protein structures, bound ligands, functional and catalytic details for each of the proteins was retrieved and carefully tabulated. Individual protein structures were analyzed and categorized as:

a) Holo and apo structures based on the presence or absence of bound ligand
b) Structures with single and multiple domains
c) Structures of protein complexes further categorized as protein–protein, protein–DNA, protein–drug and protein–peptide complexes
d) Proteins with structural motifs such as Zn finger, helix turn helix, P-loop, Greek key and Walker A and B motifs
e) Proteins expressed during dormancy[17]
f) Proteins grouped under COG[18] functional category and essential proteins required for the survival and growth of *M. tuberculosis*.
g) Structures with unknown function, as that of hypothetical proteins
h) Structures for known drug targets

2.3. Finding structural similarity of bound and cognate ligand

663 of the available 857 protein structures of *M. tb* belong to 251 proteins in complex with ligand. All ligands bound to each protein were compared using SMSD[19] software with all compounds known to be either substrates or products for the respective proteins, based on the catalytic information retrieved from SwissProt and KEGG[20] databases (MOL files for substrates and products were retrieved from KEGG database for structural comparison). The similarity between the bound ligand and the natural substrates and products of the protein is tabulated. The structural similarity is reported in terms of Tanimoto score.

2.4. Domain mapping

The amino acid sequence of all *M. tuberculosis* proteins was retrieved from SwissProt and aligned with the sequence of the different available structures for each of the proteins listed in PDB. The alignment was generated using Accelrys Discovery Studio v2.0. Mapping of the PFAM[21] domain was performed on the aligned sequences. Residues that are identical, conserved and semi conserved are colour coded for easier understanding.

2.5. Finding of structural homologues

For each protein structure having SCOP[22] (Structural Classification Of Proteins) classification, the structural homologues for each SCOP domain in the *M. tuberculosis* proteome was identified by structural superimposition using the SSM server. For protein structures that do not have SCOP classification, the entire protein structure was searched against all the available *M. tuberculosis* protein structures.

3. Results and discussion

3.1. MtbSD home page

The MtbSD home page provides a friendly interface for users. It provides important highlights of *M. tuberculosis* proteome. In the menu on the left side of the home page, users can find several links that provide information about apoprotein, holoprotein, protein complexes, proteins with structural motifs etc (Figure 1). Users can also query the database using a simple search option provided in the home page. In addition, information about the total number of genes having structures (gene index), submission form for submitting new solved structures by user and contact details (for mailing suggestions, criticisms and possible errors) are provided.

3.2. Search page

The MtbSD home page provides a simple “search option” in the menu on the left side and an “advance search” option on top of the home page. In the simple search option, the Rv number, PDB id or gene name brings the user directly to the protein information page containing all information on the queried protein. This is a simple yet faster way to search the database. With the help of the advance search option, users can filter their search by selecting different fields thereby allowing them to access specific data fulfilling the selection criteria. The complete database can be searched using keywords, protein or gene name, accession number, functional categories, etc (Figure 2). The HTML form parses the criteria in MtbSD database and returns the results in a summary table. By clicking the Rv number or MtbSD id in the summary table, detailed information on the protein will be displayed.

3.3. MtbSD statistics

MtbSD hosts information on 857 structures for 328 proteins encoded by the *M. tuberculosis* genome. Of the 857 available structures, 824 structures have been determined by X-ray crystallography, 30 structures by NMR method and 3 based on theoretical models. Statistical details for the functional categories of *M. tuberculosis* proteins are shown in Table 1.

SCOP classification is available for structures of 149 gene products. Examining the distribution of *M. tuberculosis* protein structures in SCOP database revealed that the Alpha/Beta and Alpha + Beta class of proteins were the most widely represented.