BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net ***Database***

Volume 6(3)

MTB-PCDB: *Mycobacterium tuberculosis* proteome comparison database

Lingaraja Jena, Gauri Wankhade, Satish Kumar, Bhaskar Chinnaiah Harinath*

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

Received March 09, 2011; Accepted March 16, 2011; Published April 22, 2011

Abstract:
The *Mycobacterium tuberculosis* Proteome Comparison Database (MTB-PCDB) is an online database providing integrated access to proteome sequence comparison data for five strains of *Mycobacterium tuberculosis* (H37Rv, H37Ra, CDC 1551, F11 and KZN 1435) sequenced completely so far. MTB-PCDB currently hosts 40252 protein sequence comparison data obtained through inter-strain proteome comparison of five different strains of MTB. 2373 proteins were found to be identical in all 5 strains using MTB H_{37}Rv as reference strain. To enable wide use of this data, MTB-PCDB provides a set of tools for searching, browsing, analyzing and downloading the data. By bringing together, *M. tuberculosis* proteome comparison among virulent & avirulent strains and also drug susceptible & drug resistance strains MTB-PCDB provides a unique discovery platform for comparative proteomics among these strains which may give insights into the discovery & development of TB drugs, vaccines and biomarkers.

Availability: The database is available for free at http://www.bicjbtdrc-mgims.in/MTB-PCDB/

Keywords: proteome comparison, tuberculosis, proteomic variation, virulence, drug resistance

Background:
One third of the world's population is considered to be infected with *Mycobacterium tuberculosis*, which leads to nearly 9.4 million new patients and 3 million deaths every year **[1]**. Multi-drug-resistant strains of this pathogen, emerging in association with HIV, have added a frightening dimension to the problem **[2]**. Outbreaks of extensively drug-resistant (XDR) tuberculosis have also been an increasing threat in certain regions around the world **[3].** As *M. tb* H_{37}Rv is virulent and susceptible to most of the antitubercular drugs used so far, H_{37}Ra which is an avirulent strain **[4]**, and *M. tb* KZN strain is resistant to different drugs like isoniazid, rifampicin, kanamycin, ofloxacin, ethambutol, pyrazinamide etc. **[5]**, there must be some genetic or proteomic mutations present in them. So, there is a need for genomic as well as proteomic analysis among different strains of MTB to know the variation among them. The complete genome sequences of four clinical strains of *Mycobacterium tuberculosis* (H_{37}Rv, CDC 1551, F11 and KZN 1435) and one avirulent strain H_{37}Ra is available. In this study we did proteomic comparison amongst these strains of MTB by using NCBI's standalone BLAST algorithm **[6]**.

Materials and Methodology:
Data Collection:
Whole proteome sequences of four clinical (H_{37}Rv **[7]**, CDC 1551 **[7]**, F11 **[8]** and KZN 1435 **[8]**) and one avirulent (H_{37}Ra) strains of *Mycobacterium tuberculosis* were taken from NCBI Entrez Genome database [**8**] whose complete genome sequences were available as follows: (1) *Mycobacterium tuberculosis* H_{37}Rv (GenBank version-AL123456.2, Proteins-3988). (2) *Mycobacterium tuberculosis* H_{37}Ra (GenBank version-CP000611.1, Proteins-4034). (3) *Mycobacterium tuberculosis* CDC1551 (GenBank version-AE000516.2, Proteins-4189). (4) *Mycobacterium tuberculosis* F11 (GenBank version-CP000717.1, Proteins-3941). (5) *Mycobacterium tuberculosis* KZN 1435 (GenBank version-CP001658.1, Proteins-4059).

Database Architecture & Design:
Standalone BLAST program from NCBI was also downloaded and configured for local system. The proteome sequence were formatted using formatdb program of standalone BLAST, followed by pairwise comparison (Local BLAST) among each strain using blastall program of standalone BLAST taking whole proteome at a time. *Mycobacterium tuberculosis* Proteome Comparison Database (MTB-PCDB) was developed using Microsoft SQL Server as the back end. The output of the BLAST result was then parsed and stored in MS SQL relational database tables using in-house developed PERL code. While parsing BLAST output results, percentage identities, positivities, number of gaps, identical residues, bits, bits score, e-value, query length, subject length, query sequence, subject sequence, consensus sequence etc of the first hit obtained were taken into consideration for each protein comparison.

Data Access:
The interfaces of MTB-PCDB are designed in a manner to help users in easy navigation and retrieval of information from database for analysis. The database interfaces include: Home, Statistics, Advanced Search, Advanced

Comparison, Useful Links and Help. The database can be queried to obtain the proteome sequence comparison information in different ways through a user friendly web interface as follows **(Figure 1).** The user can search protein sequence comparison data between any two strain of MTB by giving desired identities and percentage similarity. ii) Advanced Search options like identity, similarity, query coverage, bits, bits score etc. are provided for searching more specific information regarding pair wise proteome comparison. iii) A dynamic result page appears after any search in which user can sort the comparison results by identities, similarities, query coverage, bits score, query length, subject length etc. iv) The user can restrict the number of items to be shown per page obtained in searched result. v) The user can also download sequence comparison data. vi) Each comparison also navigates to the details of comparison between the two sequences of respective strains i.e., Protein Name, Protein Length, Start, End, Strand, Accession No., Gene ID, Locus, etc along with whole alignment between the query and subject sequence besides the consensus sequence showing the matches, mismatches and gaps present in the alignments between them. vii) There is also an advanced comparison page for comparing proteome of multiple strains at a time.

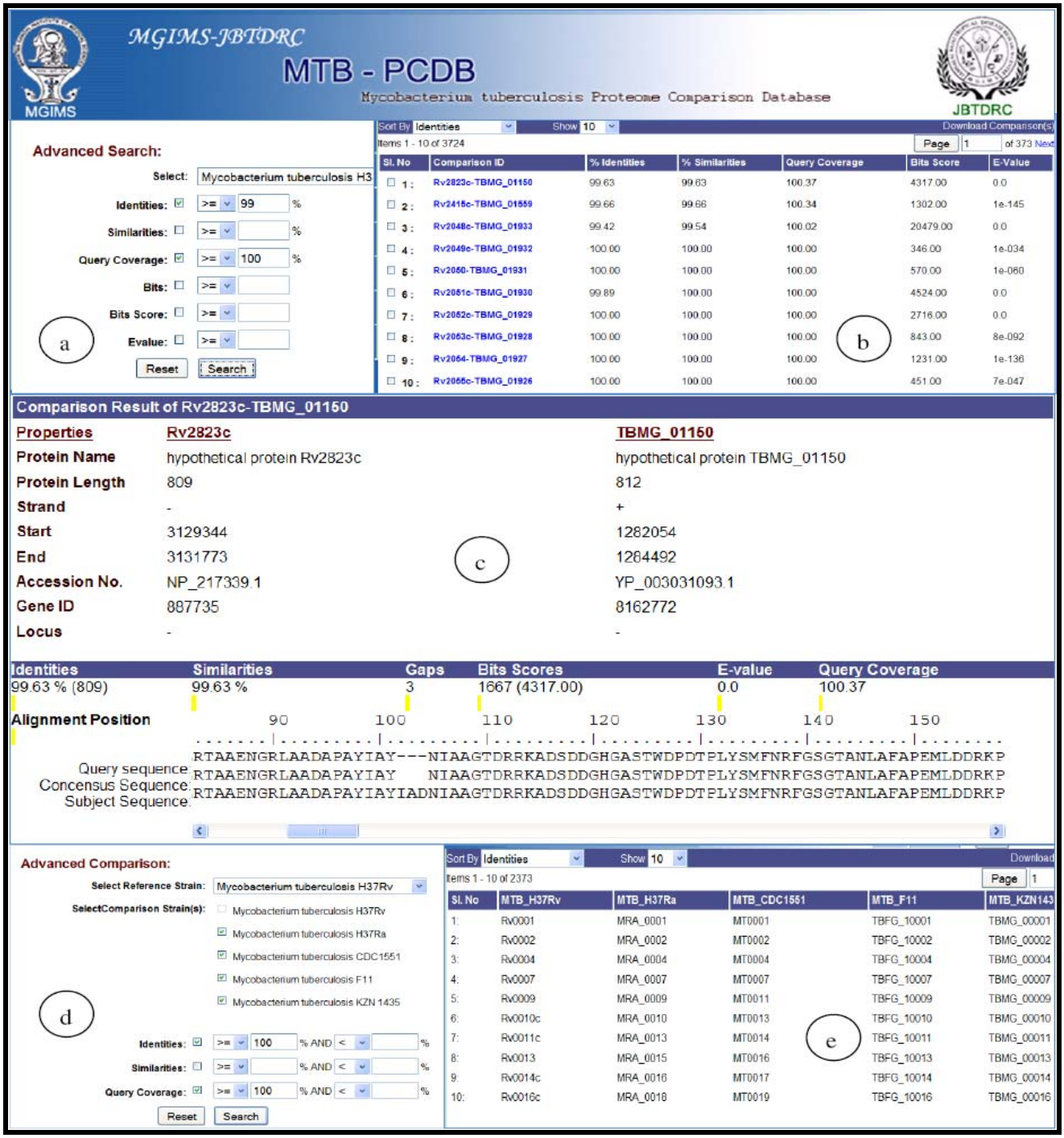

Figure 1: MTB-PCDB snapshots (a) Search option; (b) Search result; (c) Details about each comparison pair; (d) Advanced comparison; (e) Advanced comparison results

Comparison with other Databases:
Some freely available online databases also host MTB information such as Tuberculist **[9]** which includes genomic, proteomic, drugs, transcriptomics, mutant, operon annotations data etc. and Tuberculosis Database (TBDB) **[10]** provides access to genomic and annotation data of 28 different *M. tb* strains and related bacteria, also provides access to comparative genomic and microarray analysis software. For proteome comparison there is a web-based tool named Procom **[11]** for finding a subset of proteins of interest by comparing proteomes of 32 different species. However it does not consider *M. tb*. species for comparison. The uniqueness of MTB-PCDB is the involvement of comparative proteomics of five *M. tb*. strains whose genomes are completely sequenced. MTB-PCDB compares proteome of selected strains and displays detail information about each comparison pair along with the complete pair wise alignment to find out the point mutations and the consensus sequence. This may helps users to identify mutations involved in drug resistance and pathogenicity.

Utility:
MTB-PCDB, a comprehensive database with total of 40252 protein sequence comparison data. The proteomic variation found in five *M. tuberculosis* strains may have vital role in each species. This comparative study may help understand the mechanism of pathogenesis and survival of *M. tuberculosis* within the host. This information also facilitates design of new antitubercular vaccines and therapeutic agents based on the identified virulence-associated mutations.

Caveats:
MTB-PCDB does not include comparison of all the strains of *M. tb* as they are not completely sequenced.

Future Developments:
As and when in future, new TB strains are sequenced and available in public databases, we shall attempt to update MTB-PCDB including newly proteome comparison data. We would continue working on analyzing and correlating the proteomic variation among different strains with their drug resistance, virulence and pathogenic properties.

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. Technical assistance of Ms. Amrita Bit is appreciated.

References:
[1] www.who.int/tb/data
[2] Raviglione MC *et al. JAMA.* 1995 **273**: 220 [PMID: 7807661]
[3] Shah NS *et al. Emerg Infect Dis.* 2007 **13**(3): 380 [PMID: 17552090]
[4] Zheng, H *et al. PLoS One.* 2008 **3:**e2375 [PMID: 18584054]
[5] Ioerger TR *et al. PLoS One.* 2009 **4**(11):e7778. [PMID: 19890396]
[6] Altschul SF *et al. J Mol Biol.*1990 **215**: 403 [PMID: 2231712]
[7] Fleischmann RD *et al. J Bacteriol.* 2002 **184**(19): 5479 [PMID: 12218036]
[8] http://www.ncbi.nlm.nih.gov/sites/genome
[9] http://tuberculist.epfl.ch/
[10] Reddy TB *et al. Nucleic Acids Res.* 2009 **37:** 499 [PMID: 18835847].
[11] http://procom.wustl.edu/

Edited by P Kangueane
Citation: Jena *et al.* Bioinformation 6(3): 131-133 (2011)