Bioinformation
www.bioinformation.net

open access
Hypothesis

Dataset of potential targets for *Mycobacterium tuberculosis* H37Rv through comparative genome analysis

Siddiqui M. Asif[1*], Amir Asad[1], Ahmad Faizan[2], Malik S. Anjali[1], Arya Arvind[1], Kapoor Neelesh[1], Kumar Hirdesh[1], Kumar Sanjay[2]

[1]Department of Biotechnology, Meerut Institute of Engineering and Technology, Meerut; [2]Department of Biotechnology, Shobhit University, Meerut; *Corresponding author E mail: - asifsiddiqui82@gmail.com

Received November 19, 2009; Accepted December 14, 2009; Published December 31, 2009

Abstract:
Mycobacterium tuberculosis is the causative agent of the disease, tuberculosis and H37Rv is the most studied clinical strain. We use comparative genome analysis of *Mycobacterium tuberculosis* H37Rv and human for the identification of potential targets dataset. We used DEG (Database of Essential Genes) to identify essential genes in the H37Rv strain. The analysis shows that 628 of the 3989 genes in *Mycobacterium tuberculosis H37Rv* were found to be essential of which 324 genes lack similarity to the human genome. Subsequently hypothetical proteins were removed through manual curation. This further resulted in a dataset of 135 proteins with essential function and no homology to human.

Keywords: *Mycobacterium tuberculosis H37Rv*, DEG, BLASTX, targets

Background:
Mycobacterium tuberculosis (Mtb), the causative agent of tuberculosis (TB), remains a major health threat. Each year, 8 million new TB cases appear and 2 million individuals die of TB **[1].** Moreover, it is estimated that one third of the population is latently infected with Mtb, of which ~10% will develop active disease during lifetime. The development of active TB occurs when the balance between natural immunity and the pathogen changes (e.g. upon waning of protective immune response during adolescence and in HIV patients **[2]**). Further, about half a million new multi-drug resistant TB cases are estimated to occur every year **[3].** The existing drugs, although of immense value in controlling the disease to the extent that is being done today, have several shortcomings, the most important of them being the emergence of drug resistance rendering even the front-line drugs inactive. In addition, drugs such as rifampin have high levels of adverse effects making them prone for patient incompliance. Another important problem with most of the existing anti-mycobacterials, is their inability to act upon latent forms of the bacillus. In addition to these problems, the vicious interactions between the human immunodeficiency virus and TB have led to further challenges for anti-tubercular drug discovery **[4].**

The cost of research and development in the pharmaceutical industry has been rising steeply and steadily in the last decade, but the amount of time required for bringing a new product to market remains around ten to fifteen years **[5].** This problem has been labeled as an ‘‘innovation gap,’’ and it necessitates investment in inexpensive technologies that shorten the length of time spent in drug discovery. As drug discovery efforts are increasingly becoming rational and much less dependent on trial and error, identification of appropriate targets becomes a fundamental pre-requisite. As with all the other steps in drug discovery, this stage is complicated by the fact that the identified drug target must satisfy a variety of criteria to permit progression to the next stage. Important factors in this context include homology between target and host (to prevent host toxicity such homology must be low or nonexistent **[6]),** activity of the target in the diseased state **[7]** and the essentiality of the target to the pathogen’s growth and survival. Finding new targets can enhance the discovery process as well as solve the problem of drug resistance.

Traditionally, targets have been identified through established knowledge of individual protein molecules and their functions, where their function has been well-characterized. Here, we use comparative genomics for the identification of potential targets for Mtb. These methods have the advantage of speed, low cost and even more importantly, provide a systems view of the whole microbe at a time, which enables asking questions that are often difficult to address experimentally. Drug discovery has witnessed a paradigm shift from the traditional medicinal chemistry-based ligand-oriented discovery approaches to rational drug target identification and target-driven lead discovery, by targeting the molecular mechanisms of the disease.

Methodology:
Searching for the *M. tuberculosis H37Rv complete* genes
The complete genome sequence of *M. tuberculosis H37Rv* was downloaded using National Center for Biotechnology Information FTP server (www.ncbi.nlm.nih.gov/FTP).

Comparative analysis with human
The protein coding genes from *M. tuberculosis H37Rv* genome were subjected to BLAST against DEG (http://tubic.tju.edu.cn/deg) to find out the essential genes. The essential genes obtained after DEG search were compared with human genes using BLASTX. Genes which lack the homology with human were considered as potential drug target candidates for further drug development process.

Functional analysis using UNIPROT
The obtained targets genes were further analyzed by UNIPROT (www.uniprot.org) database to find out their functions.

Results:
Available data shows 3989 protein coding genes in the *M. tuberculosis* H37Rv genome. These genes were subjected to BLAST with DEG and 628 genes were found to be essential for *M. tuberculosis* H37Rv. Comparative studies with human were performed to find out genes with or without homolog to human. Genes those that were homologous to human were neglected as they were functionally similar with those of human. Out of 628 essentials genes, 324 genes lack similarity to the human genome in BLASTX homology search and were identified as potential candidates for further target based drug development. We

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(6): 245-248 (2009)

manually annotated all the genes having no homolog to human and removed hypothetical and uncharacterized genes to refine the results. The resulting dataset consist of a target dataset of 135 potential genes. These were further classified using UNIPROT based on functions. The analyzed data shows that of the 135 targets genes, 25 were involved in amino-acid biosynthesis, 10 in cell cycle, 8 in transcription, 8 in RNA Binding and 5 in Protein transport (**Table 1 in supplementary material**). It was also observed from UNIPROT results that some target genes are involved in multiple functions in different pathways (**Table 2 and Table 3 in supplementary material**).

Discussion:

Tuberculosis (TB) is a major cause of illness and death worldwide, especially in Asia and Africa. There is a death due to tuberculosis for every 15 seconds (2 million deaths per year) and about 8 million individuals develop this disease every year **[8].** Globally, 9.2 million new cases and 1.7 million deaths from TB occurred in 2006, of which 0.7 million cases and 0.2 million deaths were in HIV-positive people **[3].** The existing drugs have several shortcomings, the most important of them being the emergence of drug resistance. Existing 'front-line' anti-TB drugs include isoniazid and rifampicin. The mechanism of action of isoniazid has only become clear in the last few years [9]. Isoniazid is a pro-drug, activated intracellularly by the MTB catalase/peroxidase, probably to a radical form that irreversibly modifies the NAD(H)-binding site of one or more enzymes involved in lipid synthesis. InhA [theNADHdependent enoyl (acyl carrier protein) reductase] and KasA (the 3-oxoacyl ACP synthase) have been demonstrated to be targets for isoniazid **[9-10].** Modifications in the target enzymes and in the activating catalase/peroxidase provide the means for MTB to evade the antibiotic actions of the drug **[11-12].**

Rifampicin is a well-characterized inhibitor of DNAdependent RNA polymerases. Resistance to rifampicin results from mutations in the drug-binding site of the polymerase that do not adversely affect the enzyme's activity **[11].** Resistance to other major and 'second-line' anti-Mtb drugs is now also well known. Pyrazinamide (a close relative of nicotinamide) is also a pro-drug, requiring activation by the enzyme pyrazinamidase. Mutation in the relevant *pncA* gene affects pyrazinamide activation, and resistance may also be facilitated by alteration of its transport into the Mtb cell **[13].** Streptomycin targets translation by associating with ribosomal proteins and the 16 S RNA of the ribosome 30 S subunit. Resistance in Mtb arises from mutations in the *rspL* gene (encoding the S12 protein target for streptomycin binding) and in conserved loop regions of the 16 S RNA, encoded by the *rrs* gene **[14].** Ethambutol is known to be inhibitory to polyamine function and cell-wall synthesis. Mutations in *embB*, encoding an arabinosyltransferase involved in cell-wall biogenesis, are associated with ethambutol resistance, but other mechanisms of resistance also appear to be operative **[15].** Recently, the second-line anti-Mtb drug ethionamide has been shown to require activation by a Mtb flavin mono-oxygenase to convert it into the cytotoxic form **[16-17].** Overexpression of InhA was found to confer resistance to both isoniazid and ethionamide, revealing an obvious route for development of antibiotic resistance in the pathogen **[18].**

No new anti-Mtb drugs have been developed for well over 20 years. In view of the increasing development of resistance to the current leading anti-Mtb drugs, novel strategies are desperately needed to avert the 'global catastrophe' forecast by the WHO. The timely determination of the genome sequence of Mtb H37Rv by Stewart Cole and co-workers in 1998 provided a much-needed boost for TB research, elucidating the genetic constitution of the pathogen and revealing many novel gene products for mechanistic and structural characterization, and as potential new drug targets **[19].** Therefore, computational approach for drug targets identification, specifically for *Mtb*, can produce a list of reliable targets very rapidly. These methods have the advantage of speed, low cost and even more importantly, provide a systems view of the whole microbe at a time. Since it is generally believed that the genomes of bacteria contain both genes with and without homologues to the human host. Using computational approach for target identification is very quick to produce a desirable list. Here we performed database search and found total 3989 genes in the *M. tuberculosis H37Rv* genome, we had annotated all the genes and removed all hypothetical genes to refine the results. After removing all hypothetical genes, 135 genes have been identified as potential drug targets. These genes and their products can be targets for future drug development and even screening can be done with the available drugs for tuberculosis.

Conclusion:

Comparative genome analysis of MTB *H37Rv* and human provides a simple framework for integrating the vast amount of genomic data that can be used in the drug target identification. Drugs that specifically target genes with high homology to the host can lead to unwanted toxicity, therefore, finding new ant-ituberculosis drugs should based on subtractive genome analysis. The analysis shows that 628 of the 3989 genes in *Mycobacterium tuberculosis H37Rv* were found to be essential of which 324 genes lack similarity to the human genome. Subsequently hypothetical proteins were removed through manual curation. This further resulted in a dataset of 135 proteins with essential function and no homology to human.

References:

[1] SH Kaufmann, *Nat Rev Immunol* (2006) **6:** 699 [PMID:]
[2] P Andersen, *Trends Microbiol* (2007) **15:**7 [PMID:]
[3] World Health Organisation, *WHO report 2008*, (2008)
[4] P Nunn, *et al*. *Nat Rev Immunol*, (2005) **5:** 819
[5] http://www.roche.com/
[6] C Freiberg, *Drug Discovery Today* (2001) **6:** S72.
[7] S Wang *et al*. *Curr. Opin. Chem. Biol.* (2004) **8:**371.
[8] World Health Organization, (2003) Geneva, Switzerland.
[9] K Mdluli *et al*. *Science*, (1998) **280:** 1607
[10] DA Rozwarski *et al*. *Science*, (1998) **267:** 1638
[11] ST Cole, *Trends Microbiol*., (1994) **2:** 411
[12] B Heym *et al*. *Mol. Microbiol*., (1995) **15:** 235
[13] C Raynaud *et al*. *Microbiology*, (1999) **145:** 1359
[14] N Honore & ST Cole, *Antimicrob. Agents Chemother*., (1994) **38:** 238
[15] SV Ramaswamy *et al*. *Antimicrob. Agents Chemother*., (2000) **44:** 326
[16] TA Vanelli *et al*. *J. Biol. Chem.*, (2002) **277:** 12824
[17] AR Baulard *et al*. *J. Biol. Chem.*, (2000) **275:** 28326
[18] MH Larsen *et al*. *Mol. Microbiol.*, (2002) **46:** 453
[19] ST Cole *et al*. *Nature*, (1998) **393:** 537

Edited by P. Kangueane
Citation: Asif *et al.*, Bioinformation 4(6): 245-248 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(6): 245-248 (2009)

Table 1: Targets and UNIPROT functional assignment

No. of Targets	Function obtained using UNIPROT	No. of Targets	Function obtained using UNIPROT
25	Amino-acid biosynthesis	10	Cell cycle
08	Transcription	08	RNA binding
05	Protein transport	04	Translation
04	ATP binding	04	Transporter activity
04	Nucleotide Biosynthesis	03	Antibiotic resistance, Cell wall biogenesis/degradation
03	DNA repair	03	Electron transport chain
03	DNA replication	03	Oxido-reductase activity
03	Thiamine biosynthesis	03	Riboflavin biosynthesis
03	Transferase activity	02	Acyl transferase
02	Menaquinone biosynthesis	02	Cell wall biogenesis/degradation
02	Pantothenate biosynthesis	02	DNA binding
02	Terpenoid biosynthesis	01	Antibiotic resistance, folate biosynthesis
01	Carbohydrate metabolic process	01	Cytochrome complex assembly
01	Cobalamin biosynthesis	01	Defense response
01	Cobalamin biosynthesis, Glutamine biosynthesis	01	DNA replication, transcription
01	Fatty acid Biosynthesis, Oxidation reduction	01	Fatty acid biosynthesis
01	Folate biosynthesis	01	GTP metabolic process
01	Glutamate biosynthesis, Oxidation reduction	01	Immune response
01	Iron ion transport, Oxidation reduction	01	Lipoprotein biosynthetic process
01	Lipopolysaccharide biosynthetic process	01	Metal Binding, urease activity
01	Metal ion binding, Trehalose-phosphatase activity	01	Metal ion binding
01	Methyltransferase	01	Mo-molybdopterin cofactor biosynthetic process, Sulfur metabolic process
01	Proteolysis	01	Porphyrin biosynthesis
01	Respiratory chain complex IV assembly	01	Ribosome biogenesis
01	Virulence	01	Ribosome biogenesis, translation

Table 2: Target genes and UNIPROT data

S. No.	Gene Name	Locus	Protein ID.	Function obtained from UNIPROT
1	**embC**	Rv3793	NP_218310	Antibiotic resistance, cell wall biogenesis/degradation
2	**embB**	Rv3795	NP_218312	Antibiotic resistance, cell wall biogenesis/degradation
3	**aftA**	Rv3792	NP_218309	Antibiotic resistance, cell wall biogenesis/degradation
4	**folP1**	Rv3608c	YP_177997	Antibiotic resistance, folate biosynthesis
5	**cobQ2**	Rv3713	NP_218230	Cobalamin biosynthesis, glutamine biosynthesis
6	**dnaG**	Rv2343c	NP_216859	DNA replication, transcription
7	**des**	Rv0824c	YP_177758	Fatty acid biosynthesis, oxidation reduction
8	**gltB**	Rv3859c	NP_218376	Glutamate biosynthesis, oxidation reduction
9	**mbtG**	Rv2378c	NP_216894	Iron ion transport, oxidation reduction
10	**ureC**	Rv1850	NP_216366	Metal Binding, urease activity
11	**otsB**	Rv3372	NP_217889	Metal ion binding, trehalose-phosphatase activity
12	**moaD**	Rv3112	YP_177928	Mo-molybdopterin cofactor biosynthetic process, sulfur metabolic process
13	**rplJ**	Rv0651	NP_215165	Ribosome biogenesis, translation

Table 3: Target genes and pathways data

S. No.	Gene Name	Locus	Protein ID.	Function obtained from UNIPROT
1	embC	Rv3793	NP_218310	Cell wall biogenesis/degradation
2	embB	Rv3795	NP_218312	
3	aftA	Rv3792	NP_218309	
4	alr	Rv3423c	NP_217940	
5	pbpB	Rv2163c	NP_216679	
6	ftsQ	Rv2151c	NP_216667	Cell cycle
7	ftsZ	Rv2150c	NP_216666	
8	ftsK	Rv2748c	NP_217264	
9	murG	Rv2153c	NP_216669	
10	mraY	Rv2156c	NP_216672	
11	ftsX	Rv3101c	NP_217617	
12	xerD	Rv1701	NP_216217	
13	murC	Rv2152c	NP_216668	
14	murF	Rv2157c	NP_216673	

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(6): 245-248 (2009)

15	murD	Rv2155c	NP_216671	
16	dnaE1	Rv1547	NP_216063	DNA replication
17	dnaA	Rv0001	NP_214515	
18	dnaB	Rv0058	NP_214572	
19	dnaG	Rv2343c	NP_216859	
20	cmk	Rv1712	NP_216228	Nucleotide biosynthesis
21	pyrH	Rv2883c	NP_217399	
22	thyX	Rv2754c	NP_217270	
23	nadA	Rv1594	NP_216110	
24	folB	Rv3607c	YP_177996	Folate biosynthesis
25	folP1	Rv3608c	YP_177997	
26	rpoD	Rv2703	NP_217219	Transcription
27	mysB	Rv2710	NP_217226	
28	nusB	Rv2533c	NP_217049	
29	nusA	Rv2841c	NP_217357	
30	mce3R	Rv1963c	NP_216479	
31	rpoA	Rv3457c	NP_217974	
32	rpoZ	Rv1390	NP_215906	
33	fdxA	Rv2007c	NP_216523	Electron transport chain
34	qcrA	Rv2195	NP_216711	
35	qcrC	Rv2194	NP_216710	

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(6): 245-248 (2009)