J Mol Model (2012) 18:3993–4004
DOI 10.1007/s00894-012-1412-5

ORIGINAL PAPER

Computational structural analysis of proteins of *Mycobacterium tuberculosis* and a resource for identifying off-targets

Sameer Hassan · Abhimita Debnath · Vasantha Mahalingam · Luke Elizabeth Hanna

Received: 14 November 2011 / Accepted: 20 March 2012 / Published online: 27 April 2012

Abstract Advancement in technology has helped to solve structures of several proteins including *M. tuberculosis* (MTB) proteins. Identifying similarity between protein structures could not only yield valuable clues to their function, but can also be employed for motif finding, protein docking and off-target identification. The current study has undertaken analysis of structures of all MTB gene products with available structures was analyzed. Majority of the MTB proteins belonged to the α/β class. 23 different protein folds are used in the MTB protein structures. Of these, the TIM barrel fold was found to be highly conserved even at very low sequence identity. We identified 21 paralogs and 27 analogs of MTB based on domains and EC classification. Our analysis revealed that many of the current drug targets share structural similarity with other proteins within the MTB genome, which could probably be off-targets. Results of this analysis have been made available in the *Mycobacterium tuberculosis* Structural Database (http://bmi.icmr.org.in/mtbsd/MtbSD.php/search.php) which is a useful resource for current and novel drug targets of MTB.

S. Hassan · A. Debnath · L. E. Hanna (✉)
Department of Biomedical Informatics,
National Institute for Research in Tuberculosis,
Chetpet, Chennai, India
e-mail: hanna@trcchennai.in

V. Mahalingam
Department of Statistics,
National Institute for Research in Tuberculosis,
Chetpet, Chennai, India

L. E. Hanna
Department of Clinical Research,
National Institute for Research in Tuberculosis,
Chetpet, Chennai 600031, India

Keywords Off-targets · Proteins · Structural homologues · Structures

Introduction

It is estimated that approximately a billion people will be newly infected with tuberculosis (TB), more than 150 million people will get sick and 36 million will die of TB between 2002 and 2020 [1]. The incidence of multi-drug-resistant TB (MDR-TB), defined as resistance to rifampicin and isoniazid and possibly other drugs, is on the increase. This necessitates the identification of newer drug targets and novel small molecules for combating TB.

Sequencing of the complete genome of *Mycobacterium tuberculosis* (MTB) by Cole in 1998 [2] was a landmark achievement that opened up new avenues for mycobacterial research. It provided information on genes involved in particular cellular functions and those involved in multiple cellular functions, and the relative abundance of different gene families [3]. The genome of MTB contains 3989 genes, of which only 30 % have known functions while 70 % are still categorized as hypothetical. Analysis of hypothetical proteins reveals that no function can be inferred from their sequence alone. As the molecular function of a gene product is tightly coupled with its three dimensional structure, structural biology can play an important role in the search for and the discovery of molecular functions of these genes [4].

Less than 10 % (323/3989) of MTB gene products have three dimensional structures solved either by X-ray crystallography or NMR. The Tuberculosis Structural Genomics Consortium (TBSGC) comprised of 453 active members spread across 15 countries, has been developed with a goal of solving the structures of all the functionally significant

proteins of MTB, in order to facilitate the drug discovery process [5]. Currently, Protein Databank (PDB) contains a total of 843 structures for 323 mycobacterial gene products, indicating that more than one structure is available for certain proteins (resulting either due to mutations or complexing with multiple ligands).

It is believed that as homologous proteins evolve, the basic structure often remains more conserved than their sequence [6], such that several domains with a common precursor may no longer be detected by pairwise sequence similarity. The number of distinct structural folds is thought to be relatively small, less than 10,000 by most estimates, with many different sequences able to encode the same basic fold of the polypeptide chain [7]. Identifying similarity between proteins' structures could therefore yield valuable clues to their function, and can be employed for motif finding, phylogenetic tree reconstruction and protein docking [8]. Comparative analysis of different structures within a family could reveal the structural plasticity of the fold [9].

In this study, we have undertaken a comprehensive analysis of structures of MTB proteins, in order to understand the fold space and repertoire of folds that are most frequently used by its various proteins.

Materials and methods

PDB data set

From 843 protein structures available in PDB [10] for MTB, one representative structure for each gene product was selected resulting in a set of 358 protein structures. This set was domain delineated and not based on entire chain.

The PDB ID for the 358 protein structure data set were searched against SCOP database (1.75 release) [11] and those having SCOP classification were taken for structural analyses.

Drawing the protein structure space in MTB genome

To map the protein structure space in the genome of MTB, pairwise structural similarity analysis was performed for 488 structural domains (488×488) using MSD-SSM server [12] available at http://www.ebi.ac.uk/msd-srv/ssm/. Structural superimposition and alignment for each query structure and its hits were analyzed manually using Discovery Studio v2.0, to confirm sharing of a similar architecture.

Mapping of functionally similar structural homologues

From structural comparative analysis, structural domains that shared similar class and fold, irrespective of their sequence identity, were used to map their function using GO descriptors provided in Swissprot database.

Statistical analysis

The relationship between sequence identity and root means square deviation (RMSD) was calculated using Pearson's correlation method. The trend line was estimated by linear least square regression method. The distribution of sequence identity and RMSD for paralogues and analogues were plotted using bar and line diagram

Results

Of the 3989 MTB genes, solved protein structures are available for only 323 genes, accounting for 843 structures in PDB database, since some of the proteins have multiple experimental structures. The SCOP database was searched for structural classification of these 843 structures, and a total of 1241 SCOP entries were identified. The fraction of proteins grouped under each SCOP class is shown in (Fig. 1). We chose one protein structure for each of the 323 gene; resulting in 358 structures (more than one structure was included for a few multidomain proteins with incomplete domain structures). Only 149 of the selected structures were found to have SCOP entries (Fig. 2). The total number of SCOP entries for the 149 proteins was 488, 35 of the 149 proteins were multidomain proteins and had more than one SCOP domain, and 97 unique folds have so far been identified in MTB proteins according to SCOP database.

Structural similarity of proteins

SCOP entries for multiple chains were removed from the set of 488 entries resulting in 184 entries, and structure superimposition analysis was performed. Of the 184 SCOP domains 78 showed structural similarity. Of the 78 domains,

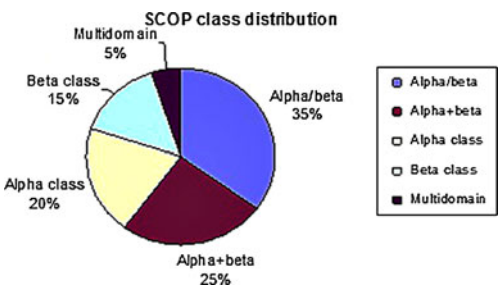

Fig. 1 Distribution of protein structures based on SCOP class. The fraction of domains in the genome of *M.tb* belonging to alpha/beta, alpha + beta, alpha and beta class are shown

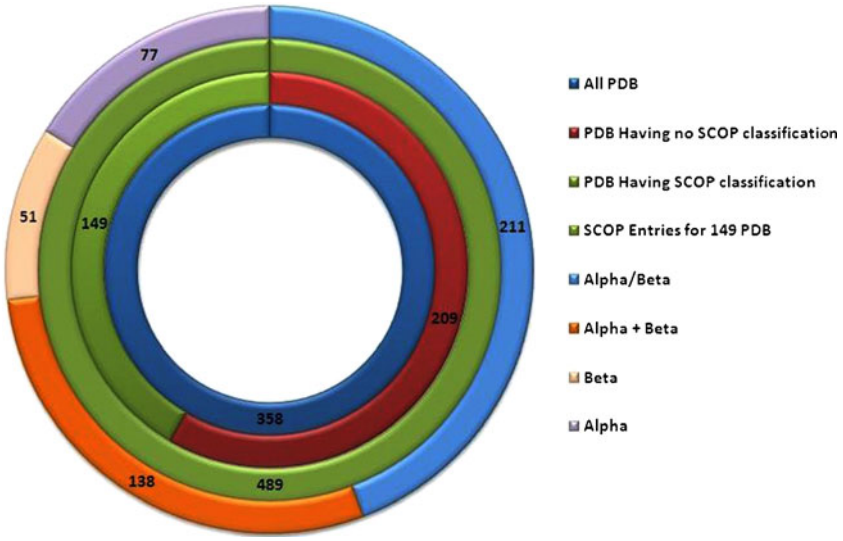

Fig. 2 SCOP distribution for the selected 358 protein structures. For the selected 358 structures, 149 structures have SCOP entries (marked as green) whereas 209 proteins are not added in the SCOP database (marked red). There are 488 SCOP entries (including multiple chains) for 149 structures (third circle colored green). The outer circle explains the fraction of each class (α/β, α + β, α, β)

65 shared structural similarity with SCOP domains and belong to the same SCOP-defined class and fold, while the remaining 13 domains shared structural architecture with SCOP domains having similar SCOP-defined class but different folds (Fig. 3). We observed that sequence identity had a significant negative correlation (−0.834, P value<0.01) with RMSD (Fig. 4). Majority of the structurally similar proteins had sequence identities ranging from 5 % to 25 % and RMSD ranging from 1.8 Å – 4.0 Å. The vast majority of the proteins in this cluster belonged to the α/β class of proteins. Using linear regression, the trend line was found to be y=3.416 - 0.039 x and 95 % confidence interval of regression coefficient (−0.045, -0.033). Its R^2 value was 0.696. The R^2 value represents a measure of its goodness-of-fit (the R^2 statistic can range from −1 to 1, with 1 representing perfect positive correlation and −1 representing perfect negative correlation).

SCOP domains sharing similar fold

Sixty five SCOP domains in the MTB proteome shared structural similarity. The sequence identity and root mean square deviation (RMSD) for the 65 SCOP domains with their hits (structural homologues) ranged from 5.7 % to 82.4 % and 0.36 Å to 4.1 Å, respectively. These 65 SCOP domains had 23 different folds, of which, the most commonly occurring fold was the TIM beta/alpha barrel (Table 1).

Major folds in the MTB proteome

TIM barrel fold

A TIM barrel fold consists of eight β/α motifs folded into a barrel structure, and is the most widely analyzed fold considering its structure, function, folding and evolution [13–17]. TIM barrel is also the most frequently used fold in MTB proteins, found even in proteins that are highly diverse at the sequence and functional level. Eight of the analyzed MTB proteins have the TIM barrel fold. All of these are enzymes having different functions, and belong to EC primary classes 2 and 4. Structure comparison revealed that five of these proteins shared structural similarity with proteins belonging to a different EC primary class. The sequence identity and RMSD between proteins having this fold ranged from 5.7 % to 16.5 %, and 2.5 Å to 3.4 Å, respectively. Though these proteins catalyze diverse reactions, their active sites are well conserved at the C-terminal end of the barrel sheet.

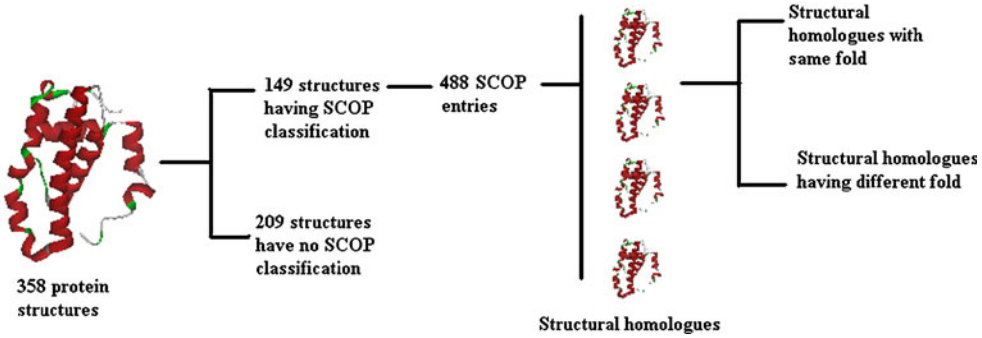

Fig. 3 A simplified flowchart representing structural comparison

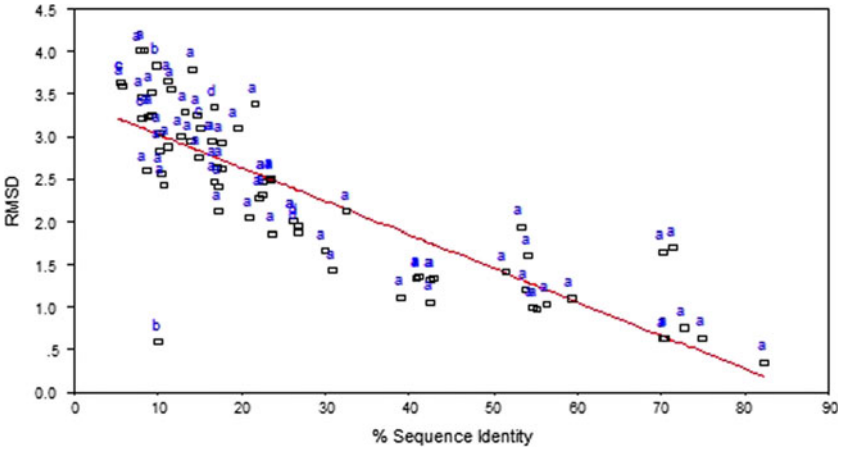

Fig. 4 Correlation between sequence identity and RMSD. The equation of the trend line is y=3.416 - 0.039 x and 95 % confidence interval of regression coefficient (−0.045, -0.033). The linear regression trend line is shown as a red line. Structures belonging to α/β (labeled as 'a'), α + β (labeled as 'b'), α (labeled as 'c') and β (labeled as 'd')

Isocitrate lyase (ICL) of MTB coded by the *icl* gene (Rv0467) plays a pivotal role in the persistence of MTB by sustaining intracellular infection in inflammatory macrophages, and is a potential drug target. This enzyme allows net carbon gain by diverting acetyl-CoA from beta-oxidation of fatty acids into the glyoxylate shunt pathway. Currently there are three experimental structures (1F61, 1F8M, 1F8I) solved for ICL covering the entire sequence. ICL belongs to the α/β class and has TIM barrel fold. Based on our analysis, we found that ketopantoate hydroxymethyltransferase (KPHMT, Rv2225), 1OY0 and dihydrodipicolinate synthase (DAPA, 1XXX, Rv2753c) of MTB share a similar fold (TIM barrel) as that of ICL (Fig. 5). The sequence identity and RMSD of 1F61 with 1OY0 and 1XXX are 16.9 % and 12.7 %, and 2.6 Å and 2.9 Å, respectively. Though the three proteins have a similar structure, they have different domain, and are involved in unrelated biological functions and processes. Chaudhuri et al. had previously identified and reported structural similarity between ICL and KPHMT and suggested that the two proteins might have arisen by divergent evolution from a common ancestor [18]. Here we report the structural similarity between ICL and DAPA.

Table 1 Fold types and number of proteins for 65 SCOP domains

Fold name[a]	Number[a]
TIM barrel	9
Ferrodoxin	2
ClpP/crotonase	2
Flavodoxin-like	4
alpha/beta-hydrolases	4
S-adenosyl-L-methionine-dependent methyltransferases	5
Ferritin-like	4
CoA-transferase family III (CaiB/BaiF)	1
P-loop containing nucleoside triphosphate hydrolases	3
Thioredoxin fold	5
NAD(P)-binding Rossmann-fold domains,	3
Split barrel-like	3
Thiolase-like	2
Chorismate mutase II	1
Thioesterase/thiol ester dehydrase-isomerase	2
Periplasmic binding protein-like II	2
ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase	1
Acyl-CoA N-acyltransferases (Nat)	2
DNA/RNA-binding 3-helical bundle	2
Globin-like	2
Cytochrome P450	2
Swivelling beta/beta/alpha domain	2
Adenine nucleotide alpha hydrolase-like	2

[a] The first column gives the name of the fold and the second column gives the number of proteins having each fold

Thioredoxin fold

Thioredoxin fold is a distinct structural motif consisting of a four-stranded β-sheet and three flanking α-helices, and is found in proteins that serve a wide variety of functions.

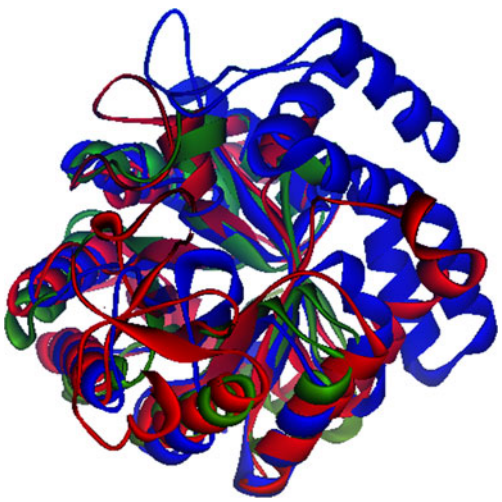

Fig. 5 Structural superimposintion of ICL (1F61,red) with 1OY0 (blue) and 1XXX (green)

Table 2 Pairwise structural comparison of proteins having AdoMet-dependent MTase fold

SCOP id	d1kp9a_	d1kpia_	d1l1ea_	d1tpya_	d2fk8a1
d1kp9a_		53.3 (1.938)	70.2 (1.656)	71.5 (1.701)	54.1 (1.612)
d1kpia_	53.3 (1.938)		59.3 (1.098)	56.4 (1.047)	51.4 (1.41)
d1l1ea_	70.2 (1.656)	59.3 (1.098)		72.8 (0.757)	54.4 (1.155)
d1tpya_	71.5 (1.701)	56.4 (1.047)	70.8 (1.111)		53.9 (1.202)
d2fk8a1	54.1 (1.612)	51.4 (1.41)	54.4 (1.155)	53.9 (1.202)	

Cells contain percentage identity and RMSD in brackets

Crystal structures are available for five proteins with this fold. The sequence identity and RMSD between these proteins range from 20 % - 55 % and 0.9 Å - 2.1 Å, respectively. Three of these five proteins, *viz.* thiol peroxidase (Rv1932, PDB code: 1XVQ), AhpC (Rv2428, PDB code: 2BMX) and AphE (Rv2238c, PDB code: 1XXU) are peroxiredoxins [19–21]. Based on the enzyme classification, Tpx, AphC and AphE belong to the oxidoreductase class of enzymes, while MPT53 (Rv2878c, PDB code: 1LU4) belongs to the class of disulfide bond-forming (Dsb) proteins [22].

Li et al. had previously identified that AphE (belonging to 1-Cys subgroup) and AphC (belonging to 2-Cys subgroup) share sequence identity of 30 % (RMSD=1.73 Å) [21]. AphE is also known to share structural similarity with AphC, Tpx and DsbF. DsbF and DsbE have a similar redoxin domain and share high sequence identity and structural similarity (55 % sequence identity and 0.98 Å RMSD).

S-adenosyl-L-methionine-dependent methyltransferases

Methyl transfers are alkylation reactions central to cellular biochemistry, and S-adenosyl-L-methionine (AdoMet) is by far the most commonly used methyl donor molecule. The AdoMet-dependent methyltransferases (MTases) act on a wide variety of target molecules, including DNA, RNA, proteins, polysaccharides, lipids and a range of small molecules. All the AdoMet-dependent MTases are reported to share a common core structure comprising of a mixture of seven stranded β sheets referred to as AdoMet-dependent MTase fold [23].

Crystal structures are available for CmaA1, CmaA2, MmaA2 and MmaA4 proteins of the eight genes that encode putative mycolic acid S-adenosylmethionine (SAM)2-dependent methyltransferases (MTs) in MTB. These proteins are responsible for catalyzing key chemical modifications in defined positions of mycolic acid [24]. They share sequence identity between 51 % and 71 %, and structural similarity between 0.75 Å and 1.5 Å (Table 2). Nine alpha helices and 7 beta strands are present in this group of proteins. Both the beta strands and alpha helices are well conserved. Among all the folds present in the proteome of MTB, AdoMet-dependent MTase fold is the most highly conserved fold with respect to sequence and structure similarity.

α/β hydrolase fold

This fold is seen in esterases, acetylcholinesterases, cutinases, carboxylesterases and epoxide hydrolases. Despite high diversity in their sequence and function, α/β-hydrolases share a common architecture and have conserved active site signatures (GxSxG and GxDxG motifs) [25].

MPT51 (PDB code: 1R88), Antigen 85A (PDB code: 1SFR), Antigen 85B (PDB code: 1F0P) and Antigen 85 C (PDB code: 1VA5) are four MTB proteins that have the canonical α/β-hydrolase fold. Antigen 85 complex and MPT51 are among the highly immunogenic secreted

Table 3 Pairwise structural comparison of proteins having the α/β hydrolase fold

SCOP id	d1r88a_	d1f0pa_	d1sfra_	d1va5a_
d1r88a_		40.9 (1.336)	42.5 (1.331)	42.9 (1.335)
d1f0pa_	40.9 (1.336)		82.4 (0.363)	75 (0.618)
d1sfra_	42.5 (1.331)	82.4 (0.363)		70.3 (0.635)
d1va5a_	42.9 (1.335)	75 (0.618)	70.3 (0.635)	

Cells contain percentage identity and RMSD in brackets

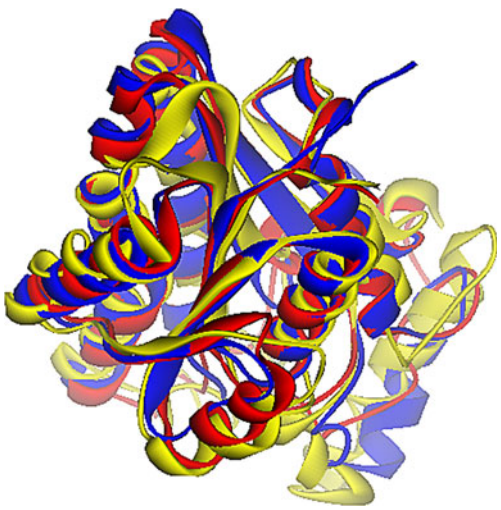

Fig. 6 Structural superimposition of *inhA* (yellow) with *fabG3* (red) and *fabG* (blue)

Table 4 Binding site details of the common patterns identified for pdb id's 1NFF (fabG3), 1UZN (fabG) and 2AQH (inhA) protein structures

Site 1: 1nff_			Site 2: 1uznA			Site 3: 2aqhA			Property	
Chain.ID	A. A.	Type	Chain.ID	A. A.	Type	Chain.ID	A. A.	Type	Source	Same AA
A.14	Gly	ACC	A.22	Gly	ACC	A.14	Gly	ACC	b	*
A.17	Arg	ALI	A.25	Arg	ALI	A.17	Thr	ALI		
A.19	Met	DON	A.27	Ile	DON	A.21	Val	DON		
A.19	Met	ALI	A.27	Ile	ALI	A.21	Val	ALI		
A.20	Gly	DON	A.28	Gly	DON	A.22	Ala	DON		
A.38	Asp	ACC	A.24	Asn	ACC	A.15	Ile	ACC		
A.60	Leu	ACC	A.60	Val	ACC	A.63	Leu	ACC		
A.61	Asp	ACC	A.61	Asp	ACC	A.64	Asp	ACC	s	*
A.61	Asp	ACC	A.61	Asp	ACC	A.64	Asp	ACC	s	*
A.62	Val	DON	A.62	Val	DON	A.65	Val	DON	b	*
A.62	Val	ALI	A.62	Val	ALI	A.65	Val	ALI	s	*
A.88	Asn	ACC	A.88	Asn	ACC	A.94	Ser	ACC		
A.88	Asn	DON	A.88	Asn	DON	A.94	Ser	DAC		
A.89	Ala	ALI	A.89	Ala	ALI	A.95	Ile	ALI		
A.90	Gly	DON	A.90	Gly	DON	A.96	Gly	DON	b	*
A.90	Gly	ACC	A.90	Gly	ACC	A.96	Gly	ACC	b	*
A.107	Arg	DON	A.107	Lys	DON	A.118	Lys	DON		
A.111	Val	ALI	A.111	Ala	ALI	A.122	Ile	ALI		
A.138	Ile	ALI	A.138	Ile	ALI	A.147	Met	ALI		
A.183	Pro	ALI	A.183	Pro	ALI	A.191	Ala	ALI		
A.186	Val	ACC	A.186	Ile	ACC	A.194	Ile	ACC		
A.186	Val	ALI	A.186	Ile	ALI	A.194	Ile	ALI		
A.188	Thr	DAC	A.188	Thr	DAC	A.196	Thr	DAC	s	*

proteins of MTB that confer pathogenicity. Though MPT51 shares significant sequence similarity with Antigen 85A, 85B and 85 C (Table 3) and has a similar esterase domain, it does not have the catalytic elements required for mycolyltransferase activity. It is therefore suggested to have non-enzymatic function and represent a new family of non-catalytic α/β-hydrolases [26]. These proteins are important drug targets for MTB.

Table 5 Proteins with different folds but sharing structural similarity

SCOP ID	Fold	Hits	Fold of hit protein	Seq. identity	RMSD
d1c3va1	NAD(P)-binding Rossmann-fold domains	d1ys6b2	Flavodoxin-like	10.1	3.038
d1gr0a1	NAD(P)-binding Rossmann-fold domains	d1rlua1	Tubulin nucleotide-binding domain-like	11.2	3.658
d1mrua_	Protein kinase-like (PK-like)	d1yk3b1	Acyl-CoA N-acyltransferases (Nat)	10	0.6
d1nkta3	P-loop containing nucleoside triphosphate hydrolases	d1kp9b_	S-adenosyl-L-methionine-dependent methyltransferases	9.2	3.521
d1pqwa_	NAD(P)-binding Rossmann-fold domains	d1kp9a_	S-adenosyl-L-methionine-dependent methyltransferases	13.9	2.944
d1pqwa_	NAD(P)-binding Rossmann-fold domains	d1nffa_	NAD(P)-binding Rossmann-fold domains	16.7	2.464
d1riia_	Phosphoglycerate mutase-like	d1ywfa1	(Phosphotyrosine protein) phosphatases II	11.6	3.568
d1rlua1	Tubulin nucleotide-binding domain-like	d1uzla1	NAD(P)-binding Rossmann-fold domains	13.3	3.286
d1rlua1	Tubulin nucleotide-binding domain-like	d1p44e_	NAD(P)-binding Rossmann-fold domains	8.7	3.238
d1t56a2	Tetracyclin repressor-like C-terminal domain	d2fp2a1	Chorismate mutase	5.6	3.639
d1t56a2	Tetracyclin repressor-like C-terminal domain	d2ao2a1	Chorismate mutase	5.6	3.639
d2ce3a1	ClpP/crotonase	d1ys6a2	Flavodoxin-like	7.8	4.002

Fig. 7 Comparison of structures having different SCOP folds. Structural similarity between dapB (1C3V, colored yellow) and prrA (1YS6, colored red) proteins

NAD binding rossmann fold

The NAD-binding Rossmann fold is one of the most common protein folds observed in a large number of enzyme families [27] and is believed to have evolved early [28]. This fold has a conserved double β-α-β-α-β motif, a common structural feature of many enzymes that bind NAD, NADP and related cofactors [27].

In the present study, three proteins viz. inhA, fabG3 and fabG, were identified to have the NAD-binding Rossmann fold. InhA protein catalyses the reduction of 2-transenoyl chains possessing at least 12 carbon atoms [29] and is involved in the FASII system for synthesis of mycolic acid [30]. It is a target for the anti-TB drug, isoniazid (INH) [31]. To inhibit inhA, INH needs to be activated by KatG protein and the

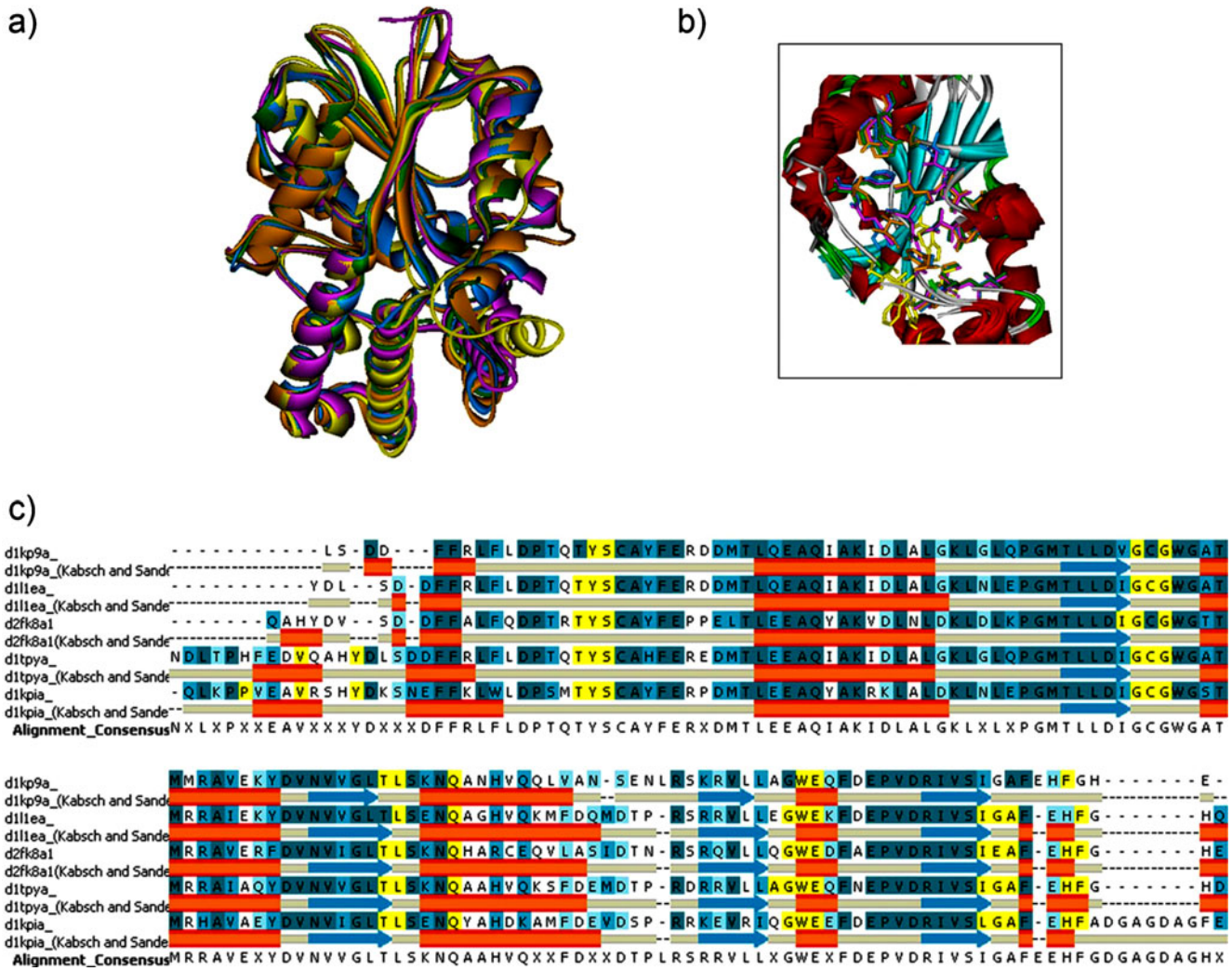

Fig. 8 Comparison of structures containing CMAS domain. **a** Superposition of CMAS domain from five Cyclopropane-fatty-acyl-phospholipid synthase proteins (2FK8-pink, 1L1E-green, 1TPY-blue, 1KP9-yellow and 1KPI- orange). **b** Superposition of the ligand binding residues from five CMAS domain. (**c**) Structure based multiple sequence alignment of the five paralogous proteins. Secondary structure is indicated by blue arrow for strands and red cylinder for helices. The ligand binding residues are shaded yellow

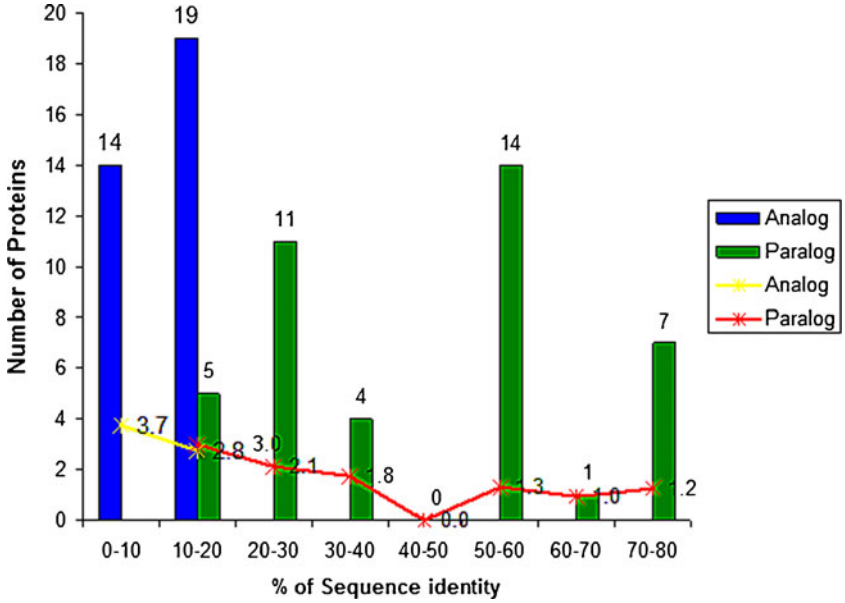

Fig. 9 Sample sizes are shown for paralogous (green bars) and analogous (blue bars) proteins for each range of sequence identities. The mean RMSD for each range of sequence identities for paralogous and analogous proteins are represented as red line and yellow line respectively

isoniazid activated intermediate forms an isonicotinyl-NAD adduct (INH-NAD) through addition of either an isonicotinic acyl anion to NAD+ or an isonicotinic acyl radical to an NAD radical. The INH-NAD adduct then binds to the active site of inhA protein [31]. Studies have shown that mutations in inhA gene or its putative promoter are responsible for INH resistance in clinical strains of MTB [32]. Currently 34 protein structures are available in PDB for inhA protein. The inhA protein has a single domain with a central core that contains a Rossmann fold supporting an NADH binding site. Our analysis revealed that fabG3 (1NFF) and fabG (1UZN) proteins share the similar fold as that of inhA (Fig. 6), although the sequence identity and RMSD between inhA and fabG3 and fabG is 23.5 % and 22.1 %, and 1.93 Å and 2.2 Å, respectively. FabG3 and fabG have similar domains (adh domain) and possess 3-alpha (or 20-beta)-hydroxysteroid dehydrogenase activity. All three proteins bind to NAD and are involved in lipid metabolism and fatty acid synthesis. Comparison of the three NAD binding sites using MultiBind [33], revealed a common spatial arrangement of physicochemical properties showing that the NAD binding region is highly conserved (Table 4). While fabG3 and fabG have similar adh_short domains, however, the domain of inhA is not known.

SCOP domains with different folds

Structural comparison of the 13 SCOP domains that had belonged to similar SCOP-defined class but different folds revealed that these proteins had structural similarity to MTB proteins belonging to different fold groups (Table 5). While the architecture of the proteins in this group is conserved, the topology is quite different. Sequence identity between these proteins ranged from 5 % to 16 %, and the RMSD ranged from 2.4 Å to 5.0 Å.

For example, the crystal structure of *DapB* gene (Rv2773c, PDB code 1C3V) encoded dihydrodipicolinate reductase (DHPR) having NAD(P)-binding Rossmann-fold domain was found to share structural similarity with the trans_reg_C domain of the transcription regulatory protein (*prrA*, Rv0903c, PDB code 1YS6) that has a Flavodoxin-like fold. *DapB* and *prrA* have a sequence identity of 10.1 % and RMSD of 3.0 Å, respectively (Fig. 7).

Paralogous and analogous proteins

Based on the functions of the selected 78 SCOP domains and their structural homologues, the proteins were grouped as paralogues and analogues. Paralogs are homologous proteins that are related by a gene duplication event, and tend to show less functional similarity than orthologs [34]. Analogs are proteins that share structural similarity but share no common features such as functional residues or unusual structural features. Such structural similarity may be due to convergence to a favorable fold [35]. Proteins sharing similar fold but belonging to different EC primary class were classified as analogues, while proteins sharing similar fold and having similar PFAM domains were categorized as paralogues.

We identified 21 paralogous protein structures, the majority of which belonged to the alpha/beta class. Their sequence identity and RMSD ranged from 14.2 % to 82.4 % and 0.3 Å to 3.7 Å respectively. Thirteen domains were identified in the paralogous proteins. Of these, cyclopropane-fatty-acyl-phospholipid synthase domain was found in five different proteins of MTB.

Enzymes that have the cyclopropane-fatty-acyl-phospholipid synthase domain catalyse the reaction: S-adenosyl-L-methionine + phospholipid olefinic fatty acid <=> S-adenosyl-L-homocysteine + phospholipid cyclopropane fatty acid. The

major mycolic acid produced by MTB contains two cis-cyclopropanes in the meromycolate chain. Cyclopropanation may contribute to the structural integrity of the cell wall complex. All five proteins identified to have the CMAS domain share high structural similarity (sequence identity=51.4 – 72.8 and RMSD=1.0 Å – 1.93 Å) (Table 2). All the secondary structural elements such as helices and strands as well as active site residues in these proteins are well conserved (Fig. 8).

We identified 27 analogous proteins. The majority of these proteins had the TIM barrel fold. Sequence identity between the analogues ranged from 8.1 to 55.1, and the RMSD ranged from 0.9 Å to 3.4 Å. In comparison with paralogous proteins, the sequence identity of analogous proteins was very low. Our findings agree with that of [34]. We plotted sequence identity versus RMSD separately for 21 pairs of paralogous proteins and 27 pairs of analogous proteins (Fig. 9). None of the paralogous proteins had sequence identity <10 % and no analogous proteins had sequence identity >30 %. Comparison of structural divergence between the two groups with similar sequence identity showed no major differences in RMSD. For example, three dimensional structures of fabG and ftsZ share good structural similarity in spite of low sequence identity (13.3 %) and no similar functional characteristics (Fig. 10). FabG3 and fabG are identified as paralogues, while ftsZ is analogous to fabG protein.

Sequence relationship between structural homologues

Sequences of all the 149 *M.tb* proteins with solved structures as well as SCOP classification were downloaded from PDB

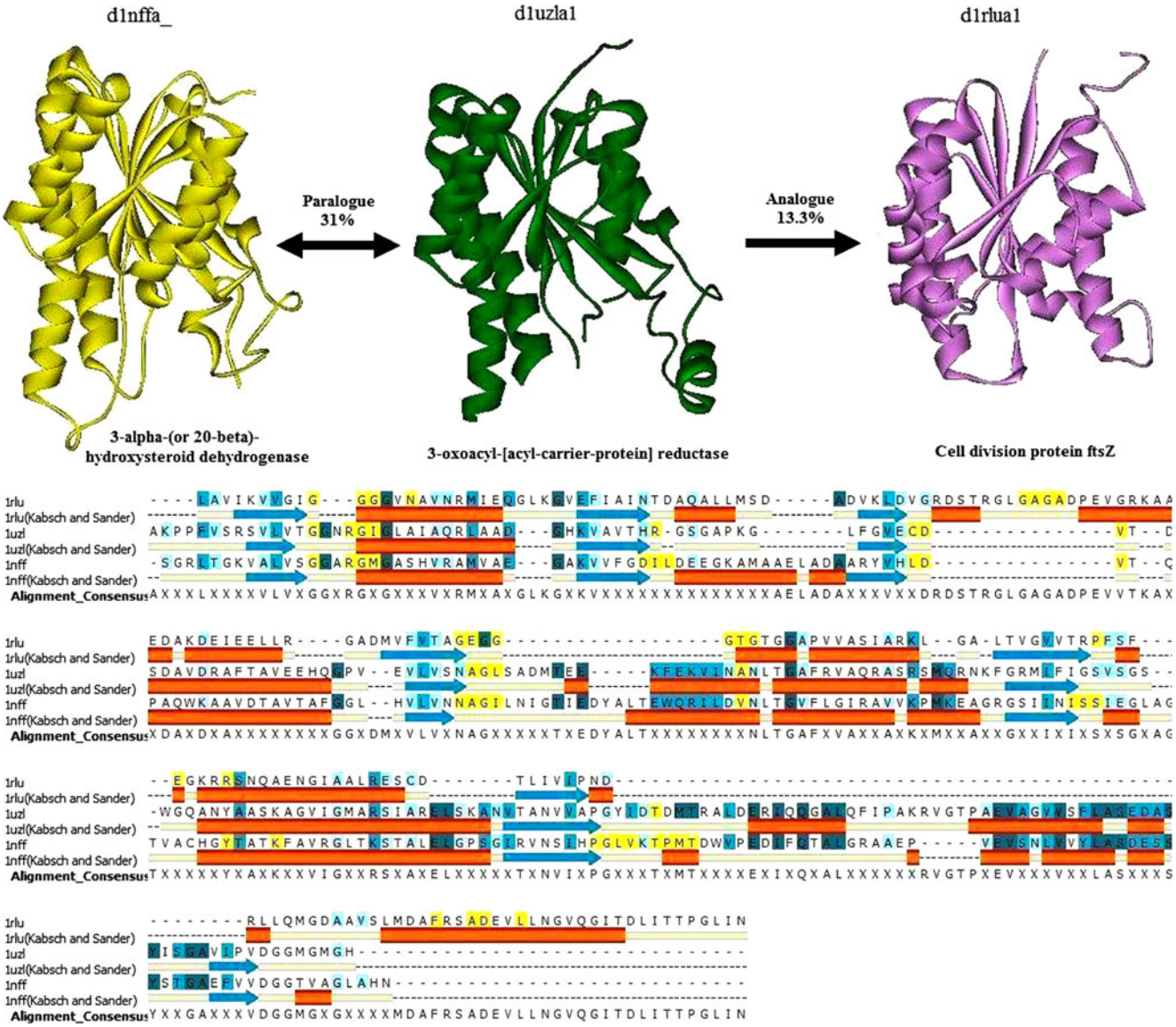

Fig. 10 **a** Schematic representation of close paralogos (PDB code: 1NFF and 1UZL) and analogous relative (PDB code: 1RLU). **b** Structure based sequence alignment of 1NFF, 1UZL and 1RLU and ligand binding site for the three proteins are marked yellow

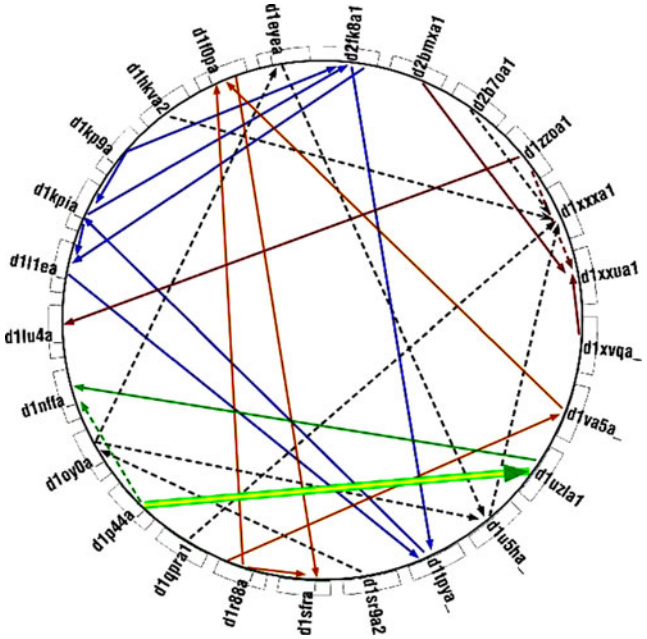

Fig. 11 SCOP wheel. Sequence relationship between protein SCOP domains belonging to five major folds in *M.tb* is shown. The different SCOP domains are labeled outside circle. SCOP domains belonging to TIM barrel fold is connected using black dotted lines. SCOP domains belonging to S- adenosyl-L-methionine-dependent proteins are connected using blue line. Domains belonging to thioredoxin fold are indicated using red line. SCOP domains belonging to Rosmann fold are connected using green line. Domains belonging to hydrolase fold are connected using orange line. Those proteins not detected by blast are shown as dotted line (e.g., TIM barrel fold), proteins which could be directly cross hit using BLAST search are shown as straight line and protein domains reported as insignificant hits using BLAST are shown as bold line (e.g.,)

and blasted against each other. Structures with sequence identity >30 % were considered as significant hits while structures with sequence identity between 20 % and 29 % were categorized as insignificant hits based on BLAST search. Structures having less than 20 % sequence identity were not detected by BLAST (Fig. 11). This exercise was undertaken to examine the degree of sequence variability between MTB proteins having similar folds.

Database for structural homologues

Drug molecules can inevitably bind not only to the intended protein target but also to other off-target proteins. There are different approaches that can be used to identify the off targets such as sequence identity between the drug target and off-target, pocket similarity, etc. In this study, we analyzed the structural similarity among proteins of MTB.

For proteins having SCOP classification, homologues are uploaded in the database as SCOP based homologues, and for protein with no SCOP classification, structural homologues were identified using domain information and uploaded as domain based homologues. This information is made available and accessible to users from Mycobacterium tuberculosis Structural Database (MtbSD) [36] and thereby making it easier for the users to identify the possible off-targets for the available protein structures of MTB.

Discussion

The first three dimensional protein structure was solved for *sodB* gene (Rv3846) of MTB in the year 1994. Availability of the full genome sequence of MTB [2] and advancement in crystallization technology has resulted in more and more protein structures being solved every year. We undertook a structural and functional comparison of MTB proteins to understand the types of folds and their frequency of usage in MTB.

Of the 358 protein structures selected in the current study 149 had SCOP classification. The majority of the proteins had SCOP folds of the α/β class; both α/β and $\alpha + \beta$ together comprised over half of the SCOP folds in the genome of MTB. This reflects the observation that the α/β class contains some of the most functionally diverse "superfolds" that act as scaffolds for a wide array of molecular and chemical functions [37]. SCOP entries for multiple chains were removed and structure superimposition analysis was performed. Of the 184 SCOP domains 78 shared structural similarity. These were further grouped into proteins having similar folds (65 proteins) and proteins having different folds (13 proteins). Most of the proteins analyzed in the present study were found to have similar folds despite statistically insignificant sequence similarity, suggesting that these folds are extra stable and have possibly diverged from a common ancestor, that despite extensive change in sequence their topology has remained the same [38].

Among structural homologues having similar fold, the TIM barrel fold was identified as the major fold. All the TIM barrel proteins of MTB function as enzymes. Sequence identity within this fold was observed to be lowest in spite of high structural similarity when compared to the other folds, such that none of these proteins could be detected using PSI-BLAST. This indicates that the TIM barrel fold is one of the most ancient and highly diverged folds. Orengo et al. reported that the TIM barrel fold is one among the nine superfolds that recur in proteins having neither sequence nor functional similarity [38]. One of the contributing factors determining the high frequency of this fold is that the kinetic folding here is straightforward compared to some of the more complex folds [38]. Thirty eight percent of the protein structures shared structural similarity with no significant

sequence and functional similarity. Knowledge of three dimensional structures not only provides unbiased structure based sequence alignment but also permits identification of conserved structural motifs not detectable by sequence analysis. Other major folds identified in MTB proteins were the S-adenosyl-L-methionine-dependent methyltransferase fold, thioredoxin fold, split barrel-like fold, P-loop containing nucleoside triphosphate hydrolases fold, flavodoxin-like and NAD (P)-binding Rossmann-fold. Except the split barrel-like fold which belongs to $\alpha + \beta$ class, all other major folds belong to α/β class.

During our analysis, we identified that inhA protein (1P44) which is an important drug target for MTB, shares structural similarity with fabG3 (1NFF) and fabG (1UZL) of MTB. All three proteins possess a similar Rossmann fold. Mutation in *inhA* gene is a major cause for resistance to INH drug. The high structural similarity between inhA, fabG3 and fabG suggests that INH-NAD adduct could bind to fabG3 and fabG proteins as well, though with varying kinetics, contributing to the development of INH resistance. FabG has previously been reported in the Tuberculosis Drug Resistance Mutation Database as being responsible for cross resistance to INH. However, here we report the structural similarity between inhA and fabG3 and suggest a probable role for this protein in cross resistance to INH. Because of the small size of the INH molecule and its binding to similar substrates such as NAD, besides an overall high degree of structural similarity, the INH-NAD adduct would be capable of forming complexes with fabG3 and fabG which in turn could lead to lower bioavailability and the efficiency of the drug, and lead to drug resistance.

Based on the structure and function, similarity and dissimilarity, around 21 proteins were identified as paralogues and 27 proteins as analogues. Active site residues in all the paralogous proteins studied were highly conserved and so also their secondary structural elements (SSEs) in spite of low sequence identity. In analogous proteins, there was high structural similarity with almost all the SSEs highly conserved in spite of insignificant sequence identity. Whereas the active sites in this group were in many cases on asimilar region of the protein structure, they shared no sequence identity with respect to the residues surrounding the active site.

Fold analysis for 149 proteins of MTB revealed that many of the proteins adopted similar folds despite low sequence identity showing that the folds are more conserved than the sequences that encode them. The list of structurally similar proteins of MTB can be accessed from MtbSD, [36]. With increasing number of protein structures being solved for MTB, new avenues that will aid in the development of new drugs are opening up.

Conclusions

This study highlights the various folds present in MTB proteins, the degree of their structural similarity and frequency of the fold usage within the MTB proteome. The most ubiquitous TIM barrel fold was also identified as the most highly diverse fold. Future work will aim at studying the active site similarity between the set of identified structurally similar proteins of MTB. From a practical perspective, understanding structural homologues within the genome will help in selecting appropriate drug targets without any off-target and designing small molecule inhibitors.

Acknowledgments The authors wish to acknowledge Indian Council of Medical Research (ICMR) - Biomedical Informatics and National Institute for Research in Tuberculosis for the funding provided. We also thank Mr. Senthilnathan, National Institute for Research in Tuberculosis (NIRT) for editing figures.

References

1. Mativandlela SP, Lall N, Meyer JJ (2008) Antibacterial, antifungal and antitubercular activity of (the roots of) Pelargonium reniforme (CURT) and Pelargonium sidoides (DC) (Geraniaceae) root extracts. S Afr J Bot 72:232–237
2. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393(6685):537–544
3. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. Nature 405(6788):823–826
4. Kim SH (2000) Structural genomics of microbes: an objective. Curr Opin Struct Biol 10(3):380–383
5. Ioerger TR, Sacchettini JC (2009) Structural genomics approach to drug discovery for Mycobacterium tuberculosis. Curr Opin Microbiol 12(3):318–325
6. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826
7. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420(6912):218–223
8. Chen L, Wu LY, Wang Y, Zhang S, Zhang XS (2006) Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison. BMC Struct Biol 6:18
9. Sierk ML, Kleywegt GJ (2004) Deja vu all over again: finding and analyzing protein structure similarities. Structure 12(12):2103–2111
10. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39 (Database issue):D392-401

11. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540
12. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D: Biol Crystallogr 60(Pt 12 Pt 1):2256–2268
13. Farber GK, Petsko GA (1990) The evolution of alpha/beta barrel enzymes. Trends Biochem Sci 15(6):228–234
14. Branden CI (1991) The TIM barrel—the most frequently occurring folding motif in proteins. Curr Opin Struct Biol 1:978–83
15. Wilmanns M, Hyde CC, Davies DR, Kirschner K, Jansonius JN (1991) Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. Biochemistry 30(38):9161–9169
16. Nagano N, Hutchinson EG, Thornton JM (1999) Barrel structures in proteins: automatic identification and classification including a sequence analysis of TIM barrels. Protein Sci 8(10):2072–2084
17. Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M (2000) Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. Science 289(5484):1546–1550
18. Chaudhuri BN, Sawaya MR, Kim CY, Waldo GS, Park MS, Terwilliger TC, Yeates TO (2003) The crystal structure of the first enzyme in the pantothenate biosynthetic pathway, ketopantoate hydroxymethyltransferase, from M tuberculosis. Structure 11(7):753–764
19. Rho BS, Hung LW, Holton JM, Vigil D, Kim SI, Park MS, Terwilliger TC, Pedelacq JD (2006) Functional and structural characterization of a thiol peroxidase from Mycobacterium tuberculosis. J Mol Biol 361(5):850–863
20. Guimaraes BG, Souchon H, Honore N, Saint-Joanis B, Brosch R, Shepard W, Cole ST, Alzari PM (2005) Structure and mechanism of the alkyl hydroperoxidase AhpC, a key element of the Mycobacterium tuberculosis defense system against oxidative stress. J Biol Chem 280(27):25735–25742
21. Li S, Peterson NA, Kim MY, Kim CY, Hung LW, Yu M, Lekin T, Segelke BW, Lott JS, Baker EN (2005) Crystal Structure of AhpE from Mycobacterium tuberculosis, a 1-Cys peroxiredoxin. J Mol Biol 346(4):1035–1046
22. Goulding CW, Apostol MI, Gleiter S, Parseghian A, Bardwell J, Gennaro M, Eisenberg D (2004) Gram-positive DsbE proteins function differently from Gram-negative DsbE homologs. a structure to function analysis of DsbE from Mycobacterium tuberculosis. J Biol Chem 279(5):3516–3524
23. Cheng X, Roberts RJ (2001) AdoMet-dependent methylation, DNA methyltransferases and base flipping. Nucleic Acids Res 29 (18):3784–3795
24. Boissier F, Bardou F, Guillet V, Uttenweiler-Joseph S, Daffe M, Quemard A, Mourey L (2006) Further insight into S-adenosylmethionine-dependent methyltransferases: structural characterization of Hma, an enzyme essential for the biosynthesis of oxygenated mycolic acids in Mycobacterium tuberculosis. J Biol Chem 281(7):4434–4445
25. Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J et al (1992) The alpha/beta hydrolase fold. Protein Eng 5(3):197–211
26. Wilson RA, Maughan WN, Kremer L, Besra GS, Futterer K (2004) The structure of Mycobacterium tuberculosis MPT51 (FbpC1) defines a new family of non-catalytic alpha/beta hydrolases. J Mol Biol 335(2):519–530
27. Lesk AM (1995) NAD-binding domains of dehydrogenases. Curr Opin Struct Biol 5(6):775–783
28. Yang S, Doolittle RF, Bourne PE (2005) Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A 102(2):373–378
29. Quemard A, Sacchettini JC, Dessen A, Vilcheze C, Bittman R, Jacobs WR Jr, Blanchard JS (1995) Enzymatic characterization of the target for isoniazid in Mycobacterium tuberculosis. Biochemistry 34(26):8235–8241
30. Marrakchi H, Laneelle G, Quemard A (2000) InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II. Microbiology 146(Pt 2):289–296
31. Oliveira JS, Pereira JH, Canduri F, Rodrigues NC, de Souza ON, de Azevedo WF Jr, Basso LA, Santos DS (2006) Crystallographic and pre-steady-state kinetics studies on binding of NADH to wild-type and isoniazid-resistant enoyl-ACP(CoA) reductase enzymes from Mycobacterium tuberculosis. J Mol Biol 359(3):646–666
32. Kapur V, Li LL, Hamrick MR, Plikaytis BB, Shinnick TM, Telenti A, Jacobs WR Jr, Banerjee A, Cole S, Yuen KY et al (1995) Rapid Mycobacterium species assignment and unambiguous identification of mutations associated with antimicrobial resistance in Mycobacterium tuberculosis by automated DNA sequencing. Arch Pathol Lab Med 119(2):131–138
33. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ (2008) MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. Nucleic Acids Res 36 (Web Server issue):W260-264
34. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 269(3):423–439
35. Sonnhammer EL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18 (12):619–620
36. Hassan S, Logambiga P, Raman AM, Subazini TK, Kumaraswami V, Hanna LE MtbSD-A comprehensive structural database for Mycobacterium tuberculosis. Tuberculosis (Edinb).
37. Orengo CA, Todd AE, Thornton JM (1999) From protein structure to function. Curr Opin Struct Biol 9(3):374–382
38. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372(6507):631–634