CORE
Metadata, citation and similar papers at co
Provided by Publications of the IAS Fellows

Enhanced functional and structural domain assignments using remote similarity detection procedures for proteins encoded in the genome of *Mycobacterium tuberculosis* H37Rv

SEEMA NAMBOORI, NATASHA MHATRE*, SENTIVEL SUJATHA**, NARAYANASWAMY SRINIVASAN† and SHASHI BHUSHAN PANDIT

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India
*Present address: *Center for Ecological Sciences, Indian Institute of Science, Bangalore 560 012, India*
***Reddy US Therapeutics, Norcross, GA 30071, USA*

†Corresponding author (Fax, 91-80-2360 0535; Email, ns@mbu.iisc.ernet.in)

The sequencing of the *Mycobacterium tuberculosis* (MTB) H37Rv genome has facilitated deeper insights into the biology of MTB, yet the functions of many MTB proteins are unknown. We have used sensitive profile-based search procedures to assign functional and structural domains to infer functions of gene products encoded in MTB. These domain assignments have been made using a compendium of sequence and structural domain families. Functions are predicted for 78% of the encoded gene products. For 69% of these, functions can be inferred by domain assignments. The functions for the rest are deduced from their homology to proteins of known function. Superfamily relationships between families of unknown and known structures have increased structural information by ~ 11%. Remote similarity detection methods have enabled domain assignments for 1325 'hypothetical proteins'. The most populated families in MTB are involved in lipid metabolism, entry and survival of the bacillus in host. Interestingly, for 353 proteins, which we refer to as MTB-specific, no homologues have been identified. Numerous, previously unannotated, hypothetical proteins have been assigned domains and some of these could perhaps be the possible chemotherapeutic targets. MTB-specific proteins might include factors responsible for virulence. Importantly, these assignments could be valuable for experimental endeavors. The detailed results are publicly available at http://hodgkin.mbu.iisc.ernet.in/~dots.

[Namboori S, Mhatre N, Sujatha S, Srinivasan N and Pandit S B 2004 Enhanced functional and structural domain assignments using remote similarity detection procedures for proteins encoded in the genome of *Mycobacterium tuberculosis* H37Rv; *J. Biosci*. **29** 245–259]

1. Introduction

The fully sequenced genomes of a number of organisms provide an opportunity to understand the molecular basis of physiology, metabolism, regulation and evolution of these organisms. Such properties are mainly inferred from the functional characterization of the gene products encoded in the genome. Computational approaches for the prediction of functional features of gene products in genomes rely on the availability of homologues that are

Keywords. Genome data analysis; hypothetical proteins, *Mycobacterium tuberculosis*; protein structures; structural and functional domains

Abbreviations used: cAMP, Adenosine 3,5′-cyclic monophosphate; cGMP, guanosine 3′,5′-cyclic monophosphate; CMAS, cyclopropane mycolic acid; cNMP, cyclic nucleotide monophosphate; DDG, 2-dehydro-3-deoxy-galactarate; HMM, hidden Markov model; LCRs, low-complexity regions; mce, mycobacterial cell entry; MS, mechanosensitive; NRDB, non-redundant database; PGAM, phosphoglycerate mutase; PSSM, position-specific scoring matrices; SDR, short chain dehydrogenase/reductases; Usp, Universal stress protein.

experimentally studied. Such prediction is of great significance in pathogenic organisms, since function recognition in these organisms can enable identification of potential drug targets. There have been several attempts, using sophisticated homology search tools, to assign functions to gene products encoded in various genomes (see for example Rychlewski *et al* 1998; Pawlowski *et al* 1999; Hoersch *et al* 2000; Tatusov *et al* 2000; Pearl *et al* 2002; Meyer *et al* 2003). Despite developments in the field of function annotation using computational methods, the process of function assignment is largely a manual, labour intensive endeavour.

Tuberculosis, caused by *Mycobacterium tuberculosis*, is well acknowledged as a chronic infectious disease widely distributed across various geographic regions and is responsible for millions of death each year. Moreover, its synergistic with HIV infection and emergence of multidrug resistant strains has made tuberculosis a global emergency (Snider *et al* 1994). Sequenced genome of *M. tuberculosis* H37Rv (Cole *et al* 1998) offers an opportunity to use various computational and experimental tools, at the genomic scale, for understanding the organism and in combating the disease. The functional prediction of the gene products encoded in *M. tuberculosis* genome is a first step towards gaining insights into the physiology of the bacterium (Cole *et al* 1998). Apart from computational prediction (Strong *et al* 2003), experimental characterization of gene products and application of methods like microarray have also contributed to the functional characterization of gene products in *M. tuberculosis* (Fisher *et al* 2002; Schroeder *et al* 2002). Since the computational prediction methods rely, mostly, on the availability of information in databases, functional prediction is an essentially ongoing process with continuous refinement of functional association of the gene products. The *M. tuberculosis* genome was annotated (Cole *et al* 1998; Müller *et al* 1999) and subsequently re-annotated (Camus *et al* 2002). These functional predictions could be used as a guiding tool in order to direct the relatively lengthier, more difficult and expensive experimental methods for exploring protein functions.

The most commonly used method for functional prediction of gene products is by identification of related well-characterized homologues using sequence-based search procedures such as BLAST (Altschul *et al* 1990) and FASTA (Pearson and Lipman 1988). But, purely sequence-based search procedures might not be able to identify proteins with low sequence similarity. However, these distantly related proteins could often be identified with the use of three-dimensional (3-D) structural information (Murzin and Bateman 1997), as the structure is usually more conserved than sequence during evolution (Chothia and Lesk 1986; Chothia and Gerstein 1997). Therefore, use of structural information could potentially enhance the functional assignments (Gerstein 1998; Huynen *et al* 1998; Hegyi and Gerstein 1999; Kelley *et al* 2000). Also, prediction of the structure with relevant biochemical motifs can provide more detailed insights into the function of proteins than sequence comparisons alone (Fischer and Eisenberg 1999; Orengo *et al* 1999; Fetrow *et al* 2001). Multiple sequence alignment of the homologues in a family is one of the methods, to obtain structurally/functionally important positions. The information in these multiple sequence alignments can be converted into position specific scoring matrices (PSSM) or profiles (Gribskov *et al* 1987). The use of profile-based search methods improves sensitivity of detection of remotely related homologues (Rychlewski *et al* 1998; Gribskov *et al* 1987; Bork and Gibson 1996; Altschul *et al* 1997; Pandit *et al* 2002). Hence, combined use of structure and profile-based method should enrich the functional assignments. The structural assignments for *M. tuberculosis* genome were attempted in the past by Müller *et al* (1999) and Buchan *et al* (2002).

Proteins can be viewed as a sequence of structural or functional domains. Similar domains can be clustered into sequence/functional families [as in Pfam (Sonnhammer *et al* 1997)] or structure families [as in SCOP (Murzin *et al* 1995)]. In our work, we have considered functional domains as the basis to infer the biological role of a protein. Since domains are relatively conserved regions in protein sequences, their identification and subsequent inference of function provides effective predictions. Moreover, domain combinations impart functional versatility to proteins. The use of BLAST search procedure in the initial step might identify homologue for only a part of the region shared between proteins, which could lead to less robust function prediction. In fact the approach followed in this paper has been applied in several early papers in context of large-scale sequence data analysis (e.g. Lewis *et al* 2000; Li *et al* 2003). The biochemical function of proteins may be inferred from the association of gene products with structural or functional domains. In the present *M. tuberculosis* H37Rv genome analysis, we have applied a combination of approaches for identifying the functional/structural domains in its gene products. The structural domains are assigned using domain families from the in-house PALI database (Balaji *et al* 2001), which is derived from SCOP. The functional domain assignments are based on families from Pfam. In the process of domain association we have applied various sensitive profile-based search methods like IMPALA (Schaffer *et al* 1999), HMMER 2.1 (Eddy 1998) and PSI-BLAST (Altschul *et al* 1997) with manual analysis of the results as the mainstay. With the present approach, it has been possible to identify a number of remotely related homologues with significant measure of reliability. The result is an enriched structural and functional prediction of the gene products encoded in *M. tuberculosis* genome. In addition, in order

to predict the gross functions of the gene products, we have drawn inference from the sequence of domains.

2. Methods

2.1 *Databases*

The multiple sequence alignments of various protein domain families were obtained from Pfam (Version 7.2) (Sonnhammer *et al* 1997), which is a database of sequence-based domain families, at the Sanger Center (http://www.sanger.ac.uk/Software/Pfam). The domain level organization of Pfam database is extremely useful, but, many similarities in the inter-domain regions, which are not deemed as domains, may go unnoticed in Pfam. This problem could be effectively addressed by following the principles as used in the construction of COG database (Tatusov *et al* 2000). In order to simplify the integration of Pfam and COG, we take the proteins without any Pfam domain assignments and search these in the non-redundant database (NRDB), thus using the basic feature of COG. The integrated structure-sequence alignment corresponding to structural families, was obtained from the in-house PALI (Release 2.1) database (http://pauling.mbu.iisc.ernet.in/~pali). The NRDB has been obtained from National Center for Biotechnology Information (ftp://ftp.ncbi.nlm.nih.gov/blast/db). The hidden Markov model (HMM) profile library was taken from Pfam (http://www.sanger.ac.uk/Software/Pfam). The *M. tuberculosis* H37Rv genome sequences were from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/).

2.2 *Sequence analysis*

We have generated IMPALA (Schaffer *et al* 1999) searchable profiles or PSSMs for the seed member alignments provided for all the families in Pfam. A family profile is generated using multiple sequence alignment, extracted from seed proteins in Pfam and one of the sequences in that family is an input for PSI-BLAST. In order to generate the profile, this 'reference' sequence is searched against a database of sequences present in the alignment input. As every hit in this search should result in picking-up of correct homologues the quality of the profile is assured. In case of PALI family profiles, multiple structure-based sequence alignments were used and PALI family sequences were enriched with homologues from Pfam or NRDB. The hmmpfam program from HMMER2.1 package (Eddy 1998) was used to match each sequence against the HMM profile library. All of these computations have been performed on linux systems and on a six-node linux cluster.

The association of gene product to functional/structural domain in each case was evaluated in terms of their statistical significance (e-value). The overview of methodology used for domain assignments is shown in figure 1. The e-value cutoff for extracting the reliable domain associations, using IMPALA, was 1×10^{-5}. This e-value cut-off has been extrapolated from the one reported by Schaffer *et al* (1999) as well as based on the benchmarking (N S Mhatre, B Anand and N Srinivasan, unpublished results) using the database of structure-based sequence alignments of similarly folded proteins. The work of Rost *et al* (2003), Devos and Valencia (2001) and Thornton (2001) cautions regarding errors in functional annotation that may arise when purely automated homology detection methods are used for assignment. Hence, in the current annotation we use a very strict e-value (1×10^{-5}) which admittedly misses several valid connections but has the advantage of avoiding false positives very effectively (S Namboori, N Srinivasan and S B Pandit, unpublished results). Moreover, our annotation procedure involves manual intervention at various stages in order to avoid, as far as possible, mistakes in homology detection. In case of HMMER hits, an e-value significant than 1×10^{-2} was used to associate domain families. For the purpose of Pfam domain assignments, results from both HMMER and IMPALA were compared. Domains recognized by HMMER were considered over IMPALA, in cases where both the search methods were able to identify domains in same gene products. The domains assigned by IMPALA were taken into consideration if (i) a gene product was uniquely assigned a domain by IMPALA but not by HMMER and (ii) a new domain was identified in an unassigned region of the gene product, which has already been associated with some domain by HMMER. For domain boundaries, the definition as per HMMER was predominantly considered. Thus, the Pfam domain assignments are a combination of HMMER and IMPALA results, as described above. The structural domain assignments, using PALI, identified in unassigned gene products were considered for merging with Pfam assignments. Importantly, the domain assignments have been manually curated so as to minimize the possibility of false-positives.

Subsequent to functional/structural domain identification, all the sequences were subjected to TMHMM2·0 (Sonnhammer *et al* 1998) in order to assign transmembrane helical regions to the gene products. Next, the sequences were queried using SEG (Wootton and Federhen 1993) program to obtain the low-complexity regions (LCRs).

The gene products with no functional/structural domain assignments were searched against NRDB using PSI-BLAST (Altschul *et al* 1997) with an e-value cut-off of 10^{-4} and threshold inclusion value of 10^{-4} for 15 rounds of iterations. The results were considered from converged round or from 15th round, whichever was earlier. At the convergence round it is ensured that the query sequence is ranked at the top of the list of hits. The other criteria used for considering hit was 60% of query coverage for

the aligned region to eliminate short insignificant regions of alignment. The NRDB search resulted in homologue identification for the gene products that could not be assigned any of functional/structural domains. Thus, even for these gene products a probable function could be suggested. Again, the PSI-BLAST results were manually curated to eliminate false positives.

We are interested in particularly analysing the hits corresponding to remote relationship with a known domain family. Those domain assignments that showed low sequence identity were analysed further. To evaluate the low sequence identity assignments, MALIGN (Johnson *et al* 1993) or hmmalign (Eddy 1998) was used to align the mycobacterial protein to its homologous sequences, extracted from Pfam or PALI. Those gene products with best pairwise sequence identity of less than or equal to 30%, were considered as remotely related domains.

3. Results and discussion

The *M. tuberculosis* H37Rv genome has 3918 gene products. We have been able to associate functions for ~ 78% of the gene products encoded in the genome. It was either

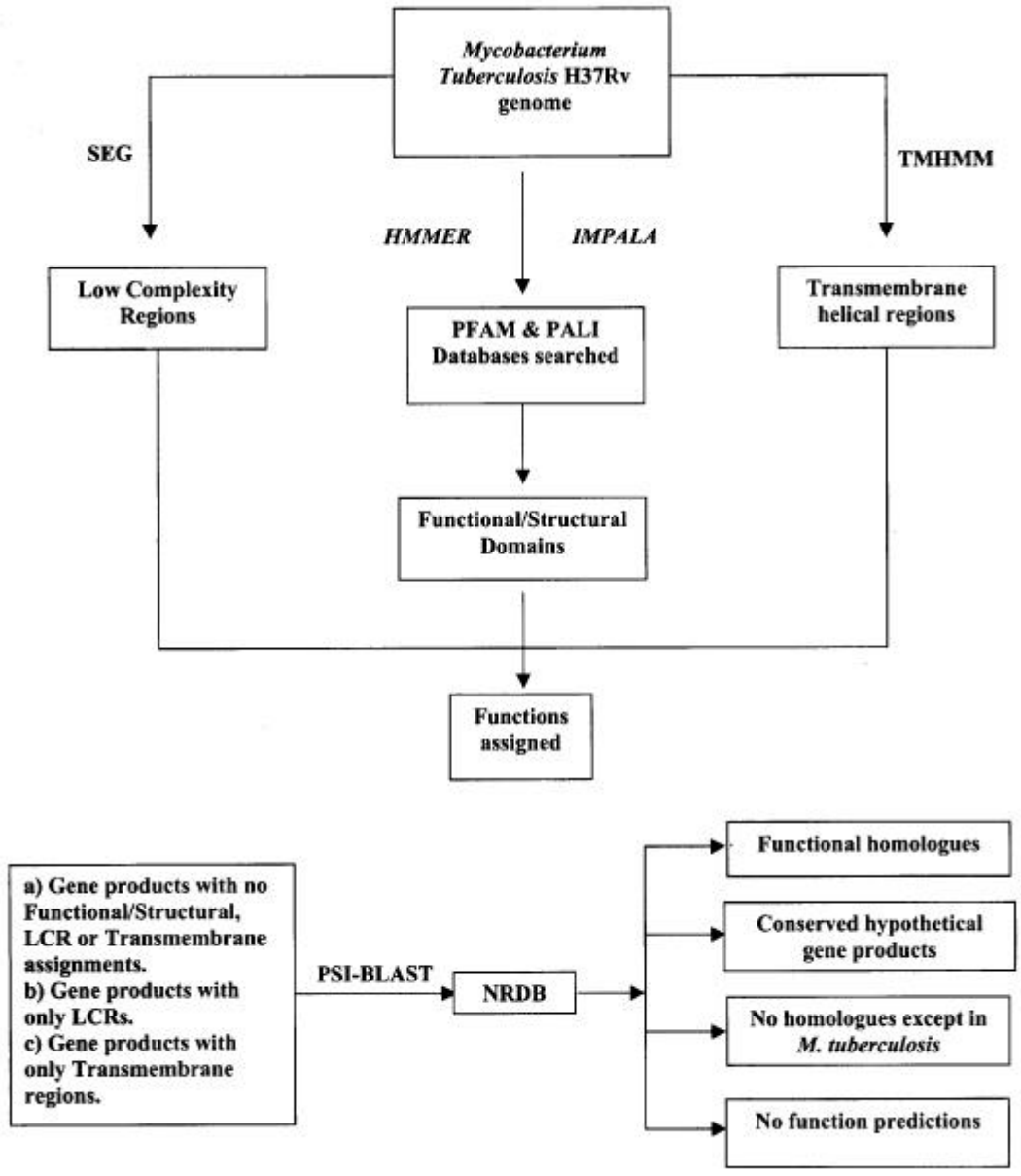

Figure 1. Overview of methodology used in the analysis.

by functional/structural domain assignment (~ 69%) or by similarity to a protein (in NRDB) of known function (~ 9%) (figure 2). Compared to the work of function association for *M. tuberculosis* protein by Camus *et al* (2002) (function annotation for 52% of the genome) and Buchan *et al* (2002) (assignment for 36% of the gene products), a higher number of gene products could be assigned function/structure in the present analysis. For ~ 9% of the gene products, search against NRDB could not identify homologues in any other organisms except in MTB itself (figure 2). About 3% of gene products were related with proteins, referred to as 'conserved hypothetical protein' from various organisms. We could not associate functions for ~ 10% of gene products since the NRDB search of these resulted in poor query coverage and/or insignificant e-values. The transmembrane and low complexity regions are identified in ~ 74% of total gene products as described in §2. Many of these occur in combination with functional/structural domains. The results of the analysis are made available at http://hodgkin.mbu.iisc.ernet.in/~dots.

3.1 *Overall statistics of genome-wide coverage of functional/structural domains*

We could assign a total of 3763 functional/structural domains to 2693 (~ 69%) proteins out of 3918 proteins encoded in the genome of *M. tuberculosis* (figure 3). The amino acids coverage by these assignments is ~ 50%. Such functional/structural domain assignments would indicate probable biochemical functions for the assigned proteins, which are useful for function prediction. Using PSI-BLAST search against NRDB for the rest of gene products, we could associate probable functions for about ~ 9%. Thus, a possible function could be suggested for a total of 3049 proteins (figure 3).

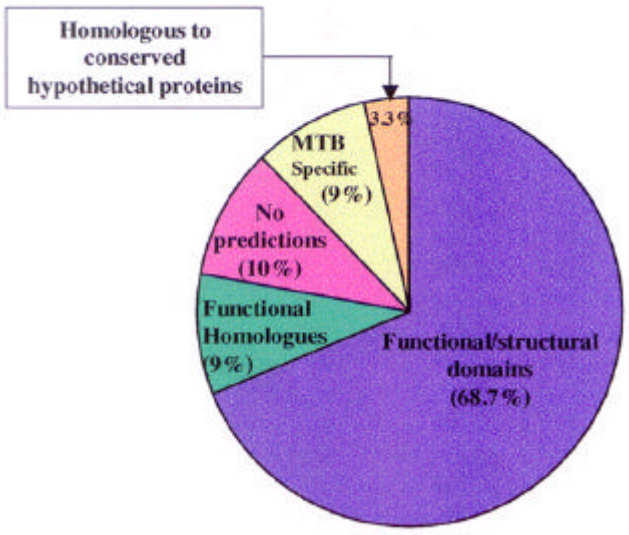

Figure 2. Pie-chart showing the overall percentage distribution of assignments for *M. tuberculosis* gene products.

The association of gene products with structure can give valuable insights, since structural information provides molecular detail of the function of a protein. The structural domain assignment will also help in prioritizing the target for TB structural genomics consortium by indicating gene products with no structural predictions. A total of 3763 domains associated with gene products belong to 1141 domain families of Pfam or PALI. Out of 1141 domain families 76 of the structural (PALI) domain families could be identified only by using PALI profiles. These 76 PALI families are present in 124 gene products with 128 domains assigned to them. Thus, use of PALI profiles has helped in enriching the domain assignments.

With a view of extending the structural information for assigned domains, we tried relating families with unknown structures to known structural families, as in SUPFAM database which was developed by us earlier (Pandit *et al* 2002). The SUPFAM database relates two or more homologous protein families, of either known or unknown structure, with use of structure-based sequence alignment. Integrating the relationships derived in SUPFAM we could provide structural information for an additional ~ 11% of domain families (figure 4). These family assignments would increase known structural content in the genome. A total of 1065 Pfam families are assigned in *M. tuberculosis* genome, of which 567 Pfam families have structural information present as in Pfam flat files. From the 498 Pfam families with no structural information, 121 families could be related to a family of known structure in SUPFAM. There are now 764 families (~ 67%) with structural information known directly or indirectly through relationships present in SUPFAM (figure 4). These 764 unique families with structural information are present in 2498 domains and would provide further insights into their functions.

As we were interested in identification of remotely related members with the use of profile based methods, we searched for remote homologues (sequence identity < = 30%), using procedures as discussed in §2. Interestingly, with the profile-based methods we could identify 223 domain associations as being remotely related and most of them were identified using PALI profiles. This suggests that the integrated structure-sequence based profiles (from PALI) have enhanced the domain assignments by identifying a number of remotely related gene products.

3.2 *Most common families in* M. tuberculosis

With the objective of identifying most frequently occurring functional domain family in *M. tuberculosis* we asses-

sed the frequency of occurrence of Pfam families. The ten most frequent Pfam families, with their functions, are listed in table 1.

Furthermore, we analysed these ten most frequently occurring Pfam families with respect to the *M. tuberculosis* biology. Among all the Pfam families in *M. tuberculosis*, the pentapeptide repeats family is the top most occurring family. The suggested role for this family is in binding of the mycobacterium to host cell receptors (Doran *et al* 1992). The PE and PPE families are among the next most frequently occurring Pfam families. PE and PPE names signify the characteristic Pro-Glu (positions 8, 9) and Pro-Pro-Glu (positions 8–10) motifs, respectively, in the conserved N-terminal domains of the protein (Cole 1999). Moreover, PE and PPE family members show preponderance of glycine, while PPE members are rich in asparagine. The PE family is observed to occur in combination with phosphoglycerate mutase (PGAM) and NHL repeats (figure 5a). PGAM is known to play a role in glycolysis and gluconeogenesis and NHL repeats are involved in protein-protein interaction. Some gene products also contain transmembrane helices in combination with PE domain. The PPE domain occurs either in combination with pentapeptide repeats or with transmembranous helical regions (figure 5b). Although no functional evidence is available for these two families, the presence

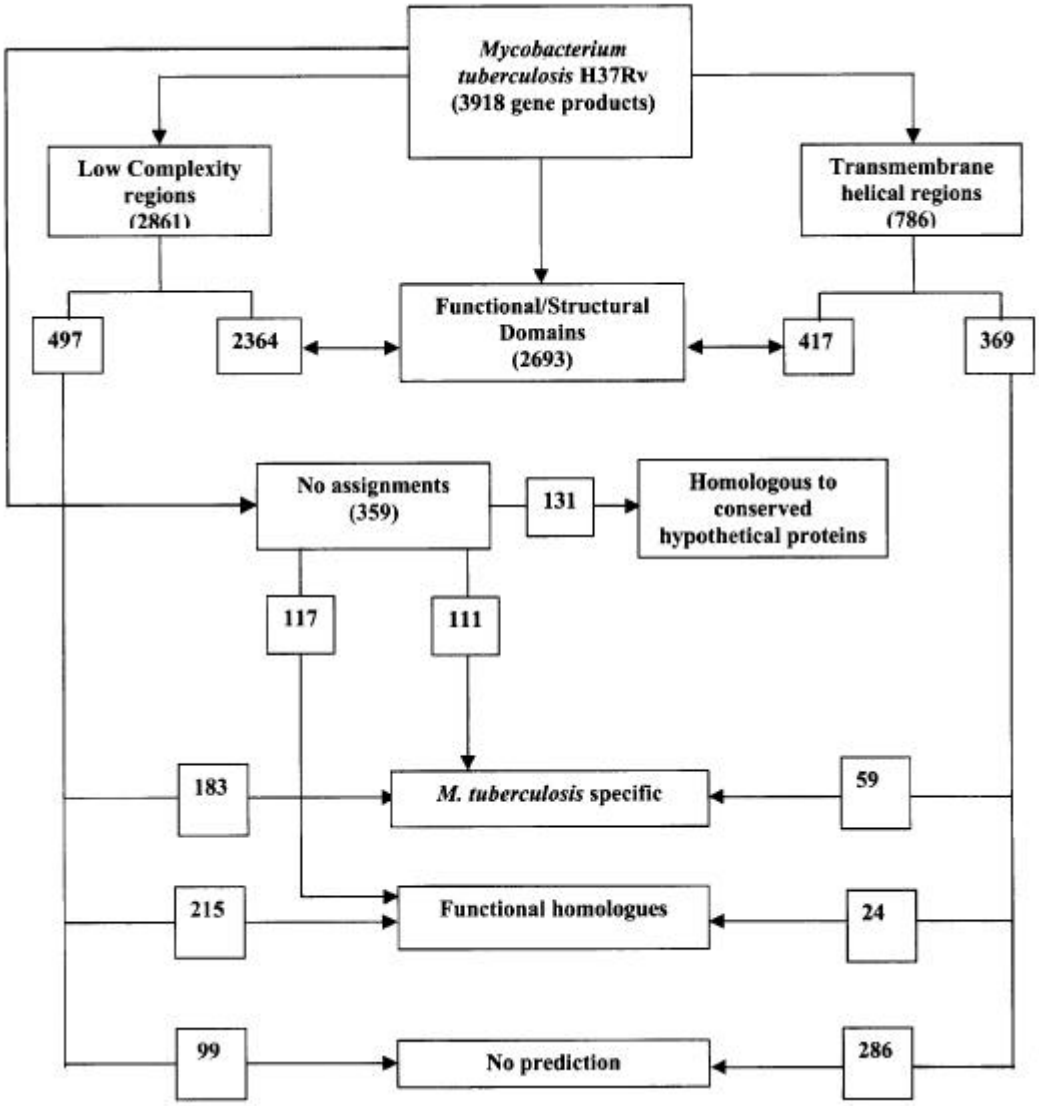

Figure 3. Schematic representation of the results showing various assignments (structural/functional domain, transmembrane and LCR) in *M. tuberculosis*. The numbers of gene products corresponding to various assignments are indicated.

of transmembrane helices indicates localization on cell surface and so, a possible role in antigenicity cannot be ruled out.

The short chain dehydrogenase/reductases (SDR) family is a member of oxidoreductase class of enzymes (Oppermann *et al* 2003). The members of this class have different substrate specificities. They are known to act on substrate such as steroids, prostaglandins, aliphatic alcohols etc. The SDR family members could be important in *M. tuberculosis* pathogenesis, driving the balance to higher levels of glucocorticoid steroids through importation and inactivation of host sex steroids. This elicits an inappropriate and ineffective T helper 2 (Th2) response as compared to a protective T helper 1 (Th1) immune response in the host (Gamieldien *et al* 2002).

The members of bacterial transcriptional regulator (tetR) family are involved in transcription regulation of a membrane-associated protein that exports the antibiotic out of the bacterial cell before it can attach to the ribosomes and inhibit polypeptide elongation (Kisker *et al* 1995). These could be involved in conferring resistance to bacterium against various antibiotics. The next commonly occurring ABC (ATP binding cassette) transporter family, found in eukaryotes and prokaryotes, constitutes

Table 1. Frequently occurring top ten Pfam families with the total number of domains observed for each.

Pfam Family	Number of domains	Function
Pentapeptide repeats	241	Binding of mycobacterium to host cell receptor
PE family	88	Highly GC-rich. Postulated role in antigenic variation and virulence of *M. tuberculosis*
PPE family	66	Highly GC-rich. Postulated role in antigenic variation and virulence of *M. tuberculosis*
Short chain dehydrogenase	61	NAD- or NADP-dependent oxidoreductases
Bacterial regulatory proteins (tetR)	52	Transcription regulator
ABC transporter	45	Translocation of compound across biological membranes
PIN	44	Signalling function
AMP-binding enzyme	44	Cellular metabolism
Acyl-CoA dehydrogenase	37	Lipid-metabolism
Abhydrolase	34	Catalytic domain found in wide range of enzymes

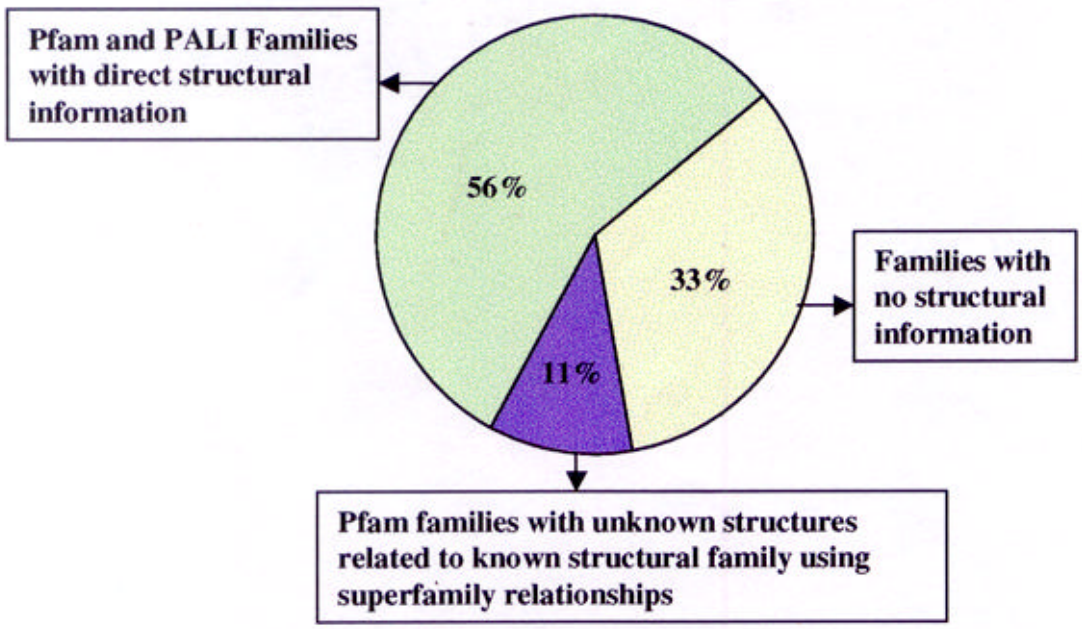

Figure 4. Pie-chart showing the distribution of Pfam and PALI families, as identified in *M. tuberculosis*, with respect to their structural information. Some families (~ 11%) with unknown structures could be associated with families of known structures using superfamily relationships.

a large superfamily of multi-subunit permeases that transports various molecules (ions, amino acids, peptides, antibiotics, polysaccharides and proteins) across biological membranes, with a relative specificity for a given substrate (Higgins 1992; Ames 1993). The mycobacterial cell wall is characterized by a highly hydrophobic cell envelope that acts as an effective permeability barrier to many compounds, hence numerous transporters would facilitate transport of various molecules. The PIN (PilT) domain family, another among the most commonly occurring families, is probably involved in a signalling function. The PIN domain occurs commonly in archaea as compared to bacteria and eukaryotes, but *Mycobacteria* seem to be an exception (Makarova *et al* 1999).

The other highly occurring families viz. AMP-binding enzyme family, acyl-CoA dehydrogenase family, abhydrolase family are involved in fatty acid and lipid metabolism. Our evaluation of the most frequently occurring families in *M. tuberculosis*, makes it apparent that gene products involved in lipid metabolism are present in sizeable numbers. This is not very surprising since *M. tuberculosis* cell envelope is known to be rich in lipids and is also one of the essential defenses for the organism. Moreover, since the cell wall structure of *M. tuberculosis* is complex, it might require a variety of transporters and permeases.

3.3 *Most abundant Pfam families in* M. tuberculosis *as compared to other prokaryotes*

We further analysed the most frequently represented Pfam families in *M. tuberculosis* by comparing the number of their occurrences with that in other prokaryotic genomes. This revealed a number of Pfam families that appeared to be most frequent in *M. tuberculosis*. A list of top fifteen families is tabulated in table 2. Of these, six families are among the top ten Pfam families, as described before. The mycobacterial cell wall is rich in lipids and hence most of the maximally occurring families are those involved in fatty acid metabolism or lipid biosynthesis. For example, Enoyl-CoA hydratase (ECH) catalyzes hydration step in fatty acid oxidation. Interestingly, one of the maximally occurring families, cytochrome p450, acts as terminal oxidase in p450-containing monooxygenase systems. They are involved in metabolism of a plethora of both exogenous and endogenous compounds such as steroids and fatty acids. These could be involved in drug resistance by degradation of drugs. The MMPL family, which is integral membrane protein, is involved in lipid transport. The abundance of lipid related enzymes indicate the significance of lipid metabolism in *M. tuberculosis* in comparison to other prokaryotes. Some of the other unusually abundant families such as mycobacterial cell entry (mce) are involved in survival of bacterium in macrophages or in virulence and antigenicity of the organism (Arruda *et al* 1993). The importance of the maximally occurring family is underlined by the fact that these are either absent or present in low numbers in many other prokaryotic genomes.

3.4 *Transmembrane and LCR assignments*

Having assigned structural/functional domain, we searched various gene products for occurrence of transmembrane helical regions using TMHMM (Sonnhammer *et al*

Table 2. Maximally occurring Pfam families in *M. tuberculosis* as compared to other prokaryotic genomes.

Pfam family	Description
PE	Highly GC-rich. Postulated role in antigenic variation and virulence of *M. tuberculosis*
PPE	Highly GC-rich. Postulated role in antigenic variation and virulence of *M. tuberculosis*
Short chain dehydrogenase	NAD- or NADP-dependent oxidoreductases
AMP-binding	Cellular metabolism
Bacterial regulatory proteins	Transcription regulator
Acyl-CoA dehydrogenase, C-terminal domain	Fatty acid biosynthesis
Acyl-CoA dehydrogenase, middle domain	Fatty acid biosynthesis
Mycobacterial cell entry related protein	Colonization and survival in macrophages
Enoyl-CoA hydratase/isomerase family	Fatty acid oxidation
Zinc-binding dehydrogenase	NADP-dependent oxidoreductase
Cytochrome P450	Metabolism of compounds like fatty acids
MMPL	Putative integral membrane proteins
O-methyltransferase N-terminus	Polyketide synthesis
Luciferase-like monooxygenase	Synthesis and degradation of various metabolites
UPF0089	Uncharacterized protein family

1998). The presence of transmembrane region suggests possible localization of gene product to the membrane. Hence, this feature could be useful in possible function prediction. The transmembrane region could be identified in 786 gene products (figure 3). Of these, 417 are found to be present in combination with functional/structural domains. This suggests that they may be localized near membranes for their function. For 369 gene products no other domain except for transmembrane regions could be assigned. The PSI-BLAST search of these against NRDB resulted in identification of functional homologues for 24 gene products. Hence, for 345 gene products we could only suggest that they are recruited at the membrane as receptors, antigens or as probable transporters, permeases etc. There were certain gene products with one transmembrane helix assigned towards N-terminus. Since it is difficult to distinguish reliably the signal peptides from transmembrane region, some of the gene products with single transmembrane segment could probably be secreted soluble proteins.

LCRs are believed to be non-globular. They may correspond to fibrous or disordered structures. Hence, characterization of LCR in gene products would also be a useful parameter. We assigned LCR to 2861 gene products (figure 3). Of these, 2364 were found to be in combination with structural/functional domains. Rest of the 497 gene products showed exclusive LCR assignments. Among the exclusively LCR assigned proteins, 215 gene products have functional homologues in NRDB, whereas 183 gene products have homologues only in *M. tuberculosis*. Since these are not known to be functionally relevant, LCRs could be given low priority in function annotation.

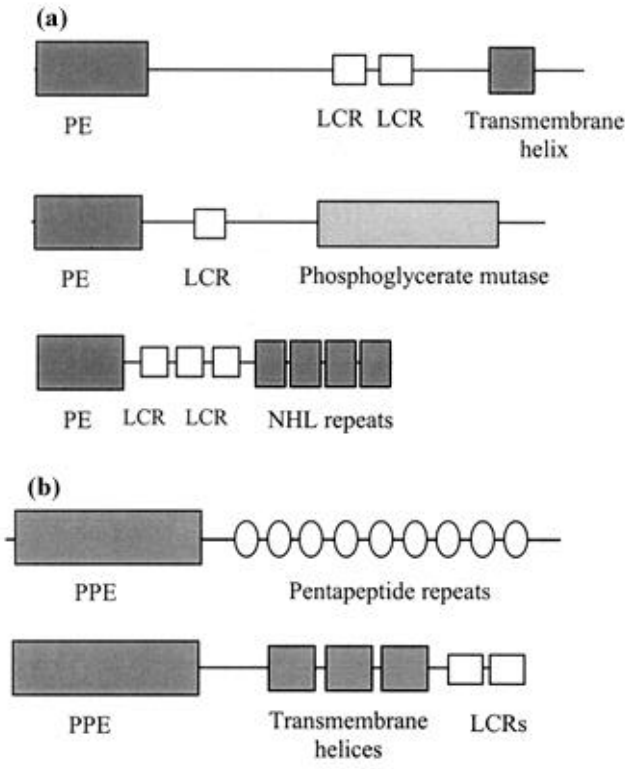

Figure 5. The domain combination observed for (**a**) PE family and (**b**) PPE family.

3.5 *Function predictions for hypothetical proteins in* M. tuberculosis

We investigated further for domain assignments in hypothetical proteins so as to attribute probable functions to these proteins. Domains were assigned for 1325 gene products, which are annotated as hypothetical proteins in the genomic database. The entire list of hypothetical proteins in the genome with domain assignments is given on our web site. Here we confine the discussion to a few of the interesting cases.

Two consecutive Universal stress protein (Usp) domains in hypothetical protein Rv2026c spanning the entire gene product were identified with e-values of $2{\cdot}9 \times 10^{-36}$ and $3{\cdot}3 \times 10^{-40}$ respectively. Our prediction concurs with those reported in databases such as SMART (Letunic *et al* 2004). The expression of Usp is enhanced several-fold during various kinds of stress like heat-shock, nutrient starvation, cell-growth inhibition by inhibitors or DNA-damaging agents (Sousa and Mckay 2001). This particular protein may be involved in similar function.

Another hypothetical protein Rv3720 could be associated with cyclopropane-fatty-acyl-phospholipid synthase family or (cyclopropane mycolic acid synthase, CMAS) domain, with an e-value of 3×10^{-76}. It could be speculated that Rv3720 might be involved in mycolic acid biosynthesis, since CMAS family members are involved in the process of lipid biosynthesis (mycolic acid synthesis). Mycolic acids are long chain *a*-alkyl-*b*-hydroxy fatty acids unique to mycobacteria. These are involved in drug resistance and survival in the hostile intracellular environment of the macrophage by the formation of an impermeable asymmetric lipid bilayer (Liu *et al* 1995). Since CMAS, which is known to occur in slow growing pathogens (George *et al* 1995), brings about mycolic acid cyclopropanation this particular hypothetical protein might play a crucial role in structural integrity of the cell-wall complex. Hence, this gene product could constitute a probable target for anti-tubercular drug-design.

Hypothetical protein Rv0170 was classified in mce protein family with an e-value of $1{\cdot}1 \times 10^{-53}$. This gene product has a signal peptide, as predicted by SignalP (http://www.cbs.dtu.dk/services/SignalP/), at its amino-terminal

that could target the protein to the host cell. *M. tuberculosis* might produce such factors, which promote its entry into the mammalian cells (Flesselles *et al* 1999). It was suggested to be important in survival and colonization of *M. tuberculosis* within macrophages (Arruda *et al* 1993). This hypothetical protein might help the tubercle bacilli to gain entry into the host macrophage.

Hypothetical protein Rv1922 was assigned a *b*-lactamase domain with an e-value of 2×10^{-08}. *b*-lactamase catalyzes the opening and hydrolysis of *b*-lactam ring of *b*-lactam antibiotics such as penicillins and thus renders these ineffective. It was shown that *M. tuberculosis* makes at least four penicillin-binding proteins (PBPs) that bind ampicillin and other *b*-lactams at clinically relevant antibiotic concentration (Chambers *et al* 1995). The outer cellular structures of tubercle bacillus do not represent any major permeability barrier for *b*-lactams (Chambers *et al* 1995; Mishra and Kasik 1970). Therefore, the production of *b*-lactamase by the bacillus seems to be the major resistance mechanism towards *b*-lactams (Voladri *et al* 1998). This hypothetical protein Rv1922 could serve as a possible target for drug-design for the design of *b*-lactamase inhibitors.

A 2-dehydro-3-deoxy-galactarate (DDG) aldolase structural domain was predicted in the hypothetical protein Rv3075c with an e-value of 8×10^{-20}. It catalyzes the reversible aldol cleavage of DDG to pyruvate and tartronic semialdehyde (Izard and Blackwell 2000). The enzyme is part of the catabolic pathway for D-glutarate/galactarate utilization in *Escherichia coli* (Hubbard *et al* 1998). Aldolases have been proven effective in biotransformations and synthesis of novel antibiotics (Wagner *et al* 1995). The catalytic residues (figure 6a) shown mapped on the 3-D structure (figure 6b), corresponding to the pdb entry 1dxe (Izard and Blackwell 2000), are conserved in hypothetical protein Rv3075c. Thus, it is possible that Rv3075c could be involved in a similar biochemical function and could serve as a probable drug target.

3.6 *Function association for proteins using domain combination*

Apart from function prediction using functional domain assignments we have used domain combination to predict the function. We have discussed few of the interesting cases below.

The hypothetical protein Rv0385 exhibits a domain combination of globin, FAD-binding and NAD-binding (figure 7a) with e-values $3{\cdot}8 \times 10^{-17}$, $5{\cdot}6 \times 10^{-31}$ and $2{\cdot}4 \times 10^{-06}$ respectively. Globin plays an essential role in binding and transport of oxygen. FAD functions as the electron carrier from $NADPH_2$ to the ferric heme prosthetic group. The flavohaemoglobin family, with similar domain architecture, includes proteins such as O_2-carrying haemoglobins and associated flavin-containing methemoglobin reductases (Zhu and Riggs 1992; Hardison 1996). One of the functions ascribed to bacterial flavohaemoglobin is nitric-oxide detoxification (Gardner *et al* 1998). Occurrence of such a domain combination in Rv0385 of *M. tuberculosis* suggests that this protein might play an important role in survival of bacterium by protecting it against oxidative damage.

Interestingly, Rv2434c has three transmembrane-spanning regions followed by mechanosensitive (MS) ion channel domain (at an e-value of $3{\cdot}1 \times 10^{-58}$) and a cyclic nucleotide monophosphate (cNMP) binding domain (at an e-value of $1{\cdot}3 \times 10^{-18}$) as shown in figure 7b. This agrees with the data reported by SMART (Letunic *et al* 2004). The MS ion channels have ability to transduce mechanical strain into electrochemical response enabling cells to respond to variety of mechanical stimuli (Chang *et al* 1998; Martinac and Kloda 2003). Cyclic NMP binding domains are present in various signal transducing proteins such as kinases, wherein they are involved in binding of cyclic nucleotides such as adenosine 3,5′-cyclic monophosphate (cAMP) and guanosine 3′,5′-cyclic monophosphate (cGMP) that act as second messenger in signalling pathways. The ion channels along with cNMP binding domain, referred to as cyclic-nucleotide gated channels, are present in eukaryotes (Finn *et al* 1996). The binding of cyclic-nucleotide in these ion channels is known to either regulate or modulate the activity of them (Finn *et al* 1996). The combination of MS channel domain and cNMP binding domain suggests that Rv2434c is a probable ion channel with dual regulation, either by mechanical stimuli or by cNMP binding. *M. tuberculosis* genome is known to possess a large number of putative cyclases (McCue *et al* 2000), which suggests a role for cNMP in a variety of signalling events and hence the modulation/regulation of MS ion channel by them is quite likely.

In Rv1318c we could assign five transmembrane regions followed by HAMP domain (at an e-value of $9{\cdot}7 \times 10^{-11}$) and adenylyl/guanylyl cyclase domain (at an e-value of $6{\cdot}3 \times 10^{-16}$) as shown in figure 7c. The HAMP domain is found in histidine kinases, methyl-accepting proteins, adenylate cyclases and other prokaryotic signalling proteins. It is probably involved in regulating the phosphorylation or methylation of homodimeric receptors by transmitting conformational changes in periplasmic domains to cytoplasmic signalling kinase and methyl-acceptor domains (Aravind and Ponting 1999). The adenylate cyclase and guanylate cyclase members in the family catalyze the formation of cAMP and cGMP respectively, which act as second messengers in downstream signalling events. The presence of HAMP domain with cyclase suggests a similar regulation of cyclase activity by HAMP domain in Rv1318c.

3.7 M. tuberculosis-*specific proteins*

We were left with 359 completely unassigned gene products for which no functional/structural domain, transmembrane and LCR assignments could be made (figure 3). These were subjected to PSI-BLAST analysis against NRDB. We could find functional homologues for 117 gene products and could suggest a possible function for them. Another 131 proteins have homologues termed as 'conserved hypothetical proteins' from other organisms. The function for these could be derived, once the experimental characterization of at least one of the homologues is available.

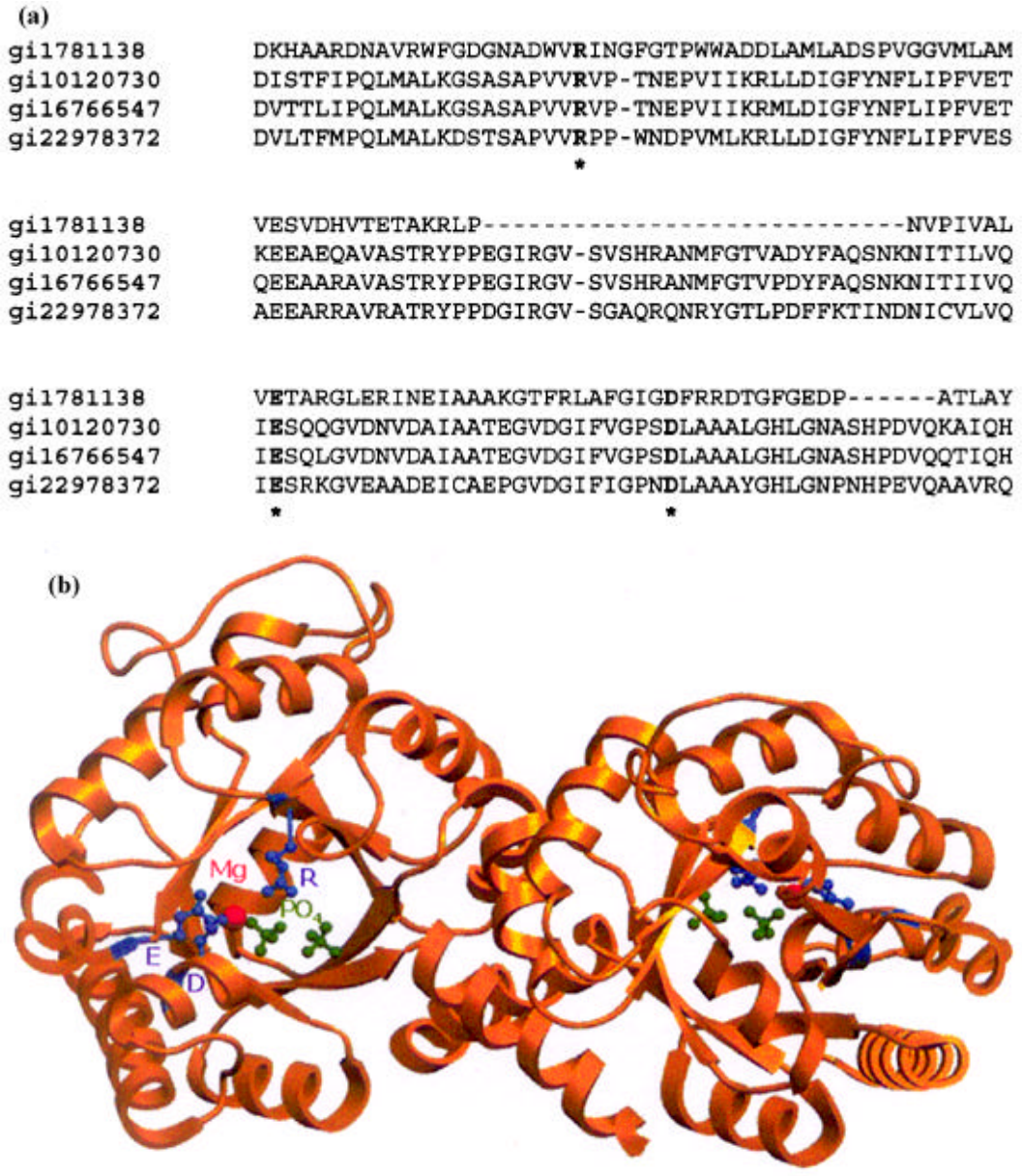

Figure 6. (**a**) Multiple sequence alignment of *M. tuberculosis* hypothetical protein Rv3075c (gi1781138), and its homologues from *E. coli* (gi10120730), *S. typhimurium* (gi16766547) and *R. metallidurans* (gi 22978372). The functional residues are highlighted in bold and marked by an asterix (*) underneath. (**b**) The 3-dimensional fold of 2-dehydro-3-deoxy-galactarate (DDG) aldolase (Izard and Blackwell 2000) which is predicted to be homologous to hypothetical protein Rv3075c. The side-chains of critical residues of DDG aldolase, which are conserved in hypothetical protein Rv3075c are shown. The figure has been produced using Setor (Evans 1993).

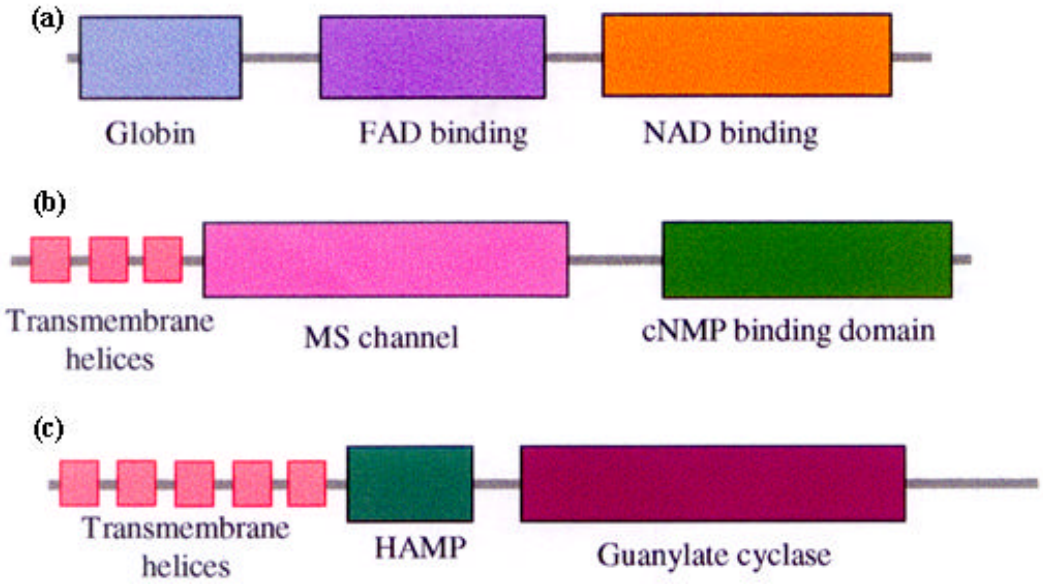

Figure 7. (**a**) The domain organization shown for hypothetical protein Rv0385. (**b**) The domain architecture for hypothetical protein Rv2434c. (**c**) The domain combination for hypothetical protein Rv1318c.

Interestingly, 111 gene products showed no homologues except in *M. tuberculosis*. Another set of gene products, which were assigned transmembrane helices and LCRs alone, consisted of 242 proteins that had homologues in *M. tuberculosis*. These form the set of 353 (111 + 242) proteins, which we refer to as *M. tuberculosis* specific gene products (figure 3). These might house some gene products, which might behave as virulence factors or play a role in uncharacterized pathways. Thus, prioritizing the experimental characterization of these would provide useful insights into the pathogenesis of the bacterium.

3.8 *Web resource*

M. tuberculosis functional/structural domain, transmembrane and LCR assignments are made publicly available for the use of scientific community on a web site. The data provides both Pfam and PALI domain assignments. The list of hypothetical proteins with domain assignments and *M. tuberculosis*-specific proteins are also provided. The list of proteins having functional homologues, identified using PSI-BLAST, is furnished. The web interface of 'DoTs' has search option for key words, gene identifier (gi code) or Rv number in the database. The site can be accessed at http://hodgkin.mbu.iisc.ernet.in/~dots.

4. Conclusions

The current functional/structural domain assignments have successfully assigned putative functions for ~ 69% of gene products, based on the known biochemical role of protein domain families. Exploring such structural and functional predictions for the *M. tuberculosis* genome has helped in enriching the functional information of the proteins encoded in the genome. In addition, protein domain architecture would help in inferring the probable biological role of the entire gene product. Another ~ 9% of proteins are related with homologues of known functions. Furthermore, using the relationships derived from SUPFAM we could provide structural information for an additional ~ 11% of domain families. These associations enhance the structural information of the genome. The structural domain association could help in prioritizing the target for structural genomics efforts.

The function predictions will be useful for experimental endeavours wherein work on the gene products without any associated function could be given preference. Moreover, some of the proteins might be promising targets for chemotherapeutic intervention. The gene products, which could not be associated with any function/structure (*M. tuberculosis* specific), could also form the priority list for structural or functional genomics efforts. These proteins might house some factors important for the virulence or persistence of *M. tuberculosis*. Ultimately,

any large-scale genome function/structure prediction is, essentially, a continuous process as the knowledge on protein domains, based on structural and other experimental work, accumulates. We will be in the process of continuous update of the structure and function information for the proteins of *M. tuberculosis*. The function assignment for *M. tuberculosis* is also made available on the web site, which would undergo regular updates.

Acknowledgements

SN is supported by computational genomics project sponsored by the Department of Biotechnology, New Delhi. SBP is supported by the Council of Scientific and Industrial Research (CSIR), New Delhi. SS was supported by the Wellcome Trust, London. This research is supported by the award of International Senior Fellowships in Biomedical Sciences to NS from the Wellcome Trust, London and also by the computational genomics project sponsored by the Department of Biotechnology, New Delhi.

References

Altschul S F, Gish W, Miller W, Myers E W and Lipman D J 1990 Basic local alignment search tool; *J. Mol. Biol.* **215** 403–410

Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z, Miller W and Lipman D J 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search program; *Nucleic Acids Res.* **25** 3389–3402

Ames G F 1993 Bacterial periplasmic permeases as model systems for the superfamily of traffic ATPases, including the multidrug resistance protein and the cystic fibrosis transmembrane conductance regulator; *Int. Rev. Cytol.* **137** 1–35

Aravind L and Ponting C P 1999 The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins; *FEMS Microbiol. Lett.* **176** 111–116

Arruda S, Bomfim G, Knights R, Huima-Byron T and Riley L W 1993 Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells; *Science* **261** 1454–1457

Balaji S, Sujatha S, Kumar S S C and Srinivasan N 2001 PALI-a database of Phylogeny and ALIgnment of homologous protein structures; *Nucleic Acids Res.* **29** 61–65

Bork P and Gibson T J 1996 Applying motif and profile searches; *Methods Enzymol.* **266** 162–184

Buchan D W, Shepherd A J, Lee D, Pearl F M, Rison S C, Thornton J M and Orengo C A 2002 Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database; *Genome Res.* **12** 503–514

Camus J, Pryor M J, Médigue C and Cole S T 2002 Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv; *Microbiology* **148** 2967–2973

Chambers H F, Moreau D, Yajko D, Miick C, Wagner C, Hackbarth C, Kocagoz S, Rosenberg E, Hadley W K and Nikaido H 1995 Can penicillins and other beta-lactam antibiotics be used to treat tuberculosis?; *Antimicrob. Agents Chemother.* **39** 2620–2624

Chang G, Spencer R H, Lee A T, Barclay M T and Rees D C 1998 Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel; *Science* **282** 2220–2226

Chothia C and Gerstein M 1997 Protein evolution. How far can sequences diverge?; *Nature* (*London*) **385** 579–581

Chothia C and Lesk A M 1986 The relation between the divergence of sequence and structure in proteins; *EMBO J.* **5** 823–826

Cole S T, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S V, Eiglmeier K, Gas S, Barry C E 3rd, *et al* 1998 Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence; *Nature* (*London*) **393** 537–544

Cole S T 1999 Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv; *FEBS Lett.* **452** 7–10

Devos D and Valencia A 2001 Intrinsic errors in genome annotation; *Trends Genet.* **17** 429–431

Doran T J, Hodgson A L, Davies J K and Radford A J 1992 Characterisation of a novel repetitive DNA sequence from *Mycobacterium bovis*; *FEMS Microbiol. Lett.* **75** 179–185

Eddy S R 1998 Profile hidden Markov models; *Bioinformatics* **14** 755–763

Evans S V 1993 SETOR: hardware-lighted three-dimensional solid model representations of macromolecules; *J. Mol. Graph.* **11** 134–138

Fetrow J S, Siew N, Di Gennaro J A, Martinez-Yamout M, Dyson J H and Skolnick J 2001 Genomic-scale comparison of sequence- and structure-based methods of function prediction: Does structure provide additional insight?; *Protein Sci.* **10** 1005–1014

Finn J T, Grunwald M E and Yau K W 1996 Cyclic nucleotide-gated ion channels an extended family with diverse functions; *Annu. Rev. Physiol.* **58** 395–426

Fischer D and Eisenberg D 1999 Predicting structures for genome proteins; *Curr. Opin. Struct. Biol.* **9** 208–211

Fisher M A, Plikaytis B B and Shinnick T M 2002 Microarray analysis of the *Mycobacterium tuberculosis* transcriptional response to the acidic conditions found in phagosomes; *J. Bacteriol.* **184** 4025–4032

Flesselles B, Anand N N, Remani J, Loosemore S M and Klein M H 1999 Disruption of the mycobacterial cell entry gene of *Mycobacterium bovis* BCG results in a mutant that exhibits a reduced invasiveness for epithelial cells; *FEMS Microbiol. Lett.* **177** 237–242

Gamieldien J, Ptitsyn A and Hide W 2002 Eukaryotic genes in *Mycobacterium tuberculosis* could have a role in pathogenesis and immunomodulation; *Trends Genet.* **18** 5–8

Gardner P R, Gardner A M, Martin L A and Salzman A L 1998 Nitric oxide dioxygenase: An enzymic function for flavohemoglobin; *Proc. Natl. Acad. Sci. USA* **95** 10378–10383

George K M, Yuan Y, Sherman D R and Barry C E 1995 The Biosynthesis of Cyclopropanated Mycolic Acids in *Mycobacterium tuberculosis*; *J. Biol. Chem.* **270** 27292–27298

Gerstein M 1998 How representative are the known structures of the proteins in a complete genome? A comprehensive structural census; *Fold. Des.* **3** 497–512

Gribskov M, McLachlan A D and Eisenberg D 1987 Profile analysis: detection of distantly related proteins; *Proc. Natl. Acad. Sci. USA* **84** 4355–4358

Hardison R C 1996 A brief history of hemoglobins: Plant, animal, protist, and bacteria; *Proc. Natl. Acad. Sci. USA* **93** 5675–5679

Hegyi H and Gerstein M 1999 The relationship between protein structure and function a comprehensive survey with application to the yeast genome; *J. Mol. Biol.* **288** 147–164

Higgins C F 1992 ABC transporters: From microorganisms to man; *Annu. Rev. Cell Biol.* **8** 67–113.

Hoersch S, Leroy C, Brown N P, Andrade M A and Sander C 2000 The GeneQuiz web server protein functional analysis through the Web; *Trends Biochem. Sci.* **25** 33–35

Hubbard B K, Koch M, Palmer D R, Babbitt P C and Gerlt J A 1998 Evolution of enzymatic activities in the enolase superfamily characterization of the (D)-glucarate/galactarate catabolic pathway in *Escherichia coli*; *Biochemistry* **37** 14369–14375

Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y and Bork P 1998 Homology-based fold predictions for *Mycoplasma genitalium* proteins; *J. Mol. Biol.* **280** 323–326

Izard T and Blackwell N C 2000 Crystal structures of the metal-dependent 2-dehydro-3-deoxy-galactarate aldolase suggest a novel reaction mechanism; *EMBO J.* **19** 3849–3856

Johnson M S, Overington J P and Blundell T L 1993 Alignment and searching for common protein folds using a data bank of structural templates; *J. Mol. Biol.* **231** 735–752

Kelley L A, MacCallum R M and Sternberg M J 2000 Enhanced genome annotation using structural profiles in the program 3D-PSSM; *J. Mol. Biol.* **299** 499–520

Kisker C, Hinrichs W, Tovar K, Hillen W and Saenger W 1995 The Complex Formed Between Tet Repressor and Tetracycline-Mg^{2+} Reveals Mechanism of Antibiotic Resistance; *J. Mol. Biol.* **247** 260–280

Lewis S, Ashburner M and Reese M G 2000 Annotating eukaryote genomes; *Curr. Opin. Struct. Biol.* **10** 349–354

Li W W, Quinn G B, Alexandrov N N, Bourne P E and Shindyalov I N 2003 A comparative proteomics resource: proteins of *Arabidopsis thaliana*; *Genome Biol.* **4** R51 Epub

Liu J, Rosenberg E Y and Nikaido H 1995 Fluidity of the Lipid Domain of Cell Wall From *Mycobacterium chelonae*; *Proc. Natl. Acad. Sci. USA* **92** 11254–11258

Letunic I, Copley R R, Schmidt S, Ciccarelli F D, Doerks T, Schultz J, Ponting C P and Bork P 2004 SMART 4·0: towards genomic data integration; *Nucleic Acids Res.* **32** D142–144

Makarova K S, Aravind L, Galperin M Y, Grishin N V, Tatusov R L, Wolf Y I and Koonin E V 1999 Comparative Genomics of the Archaea (Euryarchaeota) Evolution of Conserved Protein Families, the Stable Core, and the Variable Shell; *Genome Res.* **9** 608–628

Martinac B and Kloda A 2003 Evolutionary origins of mechanosensitive ion channels; *Prog. Biophys. Mol. Biol.* **82** 11–24

McCue L A, McDonough K A and Lawrence C E 2000 Functional classification of cNMP-binding proteins and nucleotide cyclases with implications for novel regulatory pathways in *Mycobacterium tuberculosis*; *Genome Res.* **10** 204–219

Meyer F, Goesmann A, McHardy A C, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, *et al* 2003 GenDB-an open source genome annotation system for prokaryote genomes; *Nucleic Acids Res.* **31** 2187–2195

Mishra R K and Kasik J E 1970 The mechanisms of mycobacterial resistance to penicillins and cephalosporins; *Int. J.Clin. Pharmacol.* **3** 73–77

Müller A, MacCallum R M and Sternberg M J E 1999 Benchmarking PSI-BLAST in Genome Annotation; *J. Mol. Biol.* **293** 1257–1271

Murzin A G and Bateman A 1997 Distant homology recognition using structural classification of proteins; *Proteins* (Suppl. 1) 105–112

Murzin A G and Brenner S E, Hubbard T and Chothia C 1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures; *J. Mol. Biol.* **247** 536–540

Oppermann U, Filling C, Hult M, Shafqat N, Wu X, Lindh M, Shafqat J, Nordling E, Kallberg Y, Personn B, *et al* 2003 Short-chain dehydrogenases/reductases (SDR): the 2002 update; *Chem. Biol. Interact.* **143–144**, 247–253

Orengo C A, Todd A E and Thornton J M 1999 From protein structure to function; *Curr. Opin. Struct. Biol.* **9** 374–382

Pandit S B, Gosar D, Abhiman S, Sujatha S, Dixit S S, Mhatre N S, Sowdhamini R and Srinivasan N 2002 SUPFAM-a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes; *Nucleic Acids Res.* **30** 289–293

Pawlowski K, Zhang B, Rychlewski L and Godzik A 1999 The *Helicobacter pylori* genome from sequence analysis to structural and functional predictions; *Proteins* **36** 20–30

Pearl F M, Lee D, Bray J E, Buchan D W, Shepherd A J and Orengo C A 2002 The CATH extended protein-family database providing structural annotations for genome sequences; *Protein Sci.* **11** 233–244

Pearson W R and Lipman D J 1988 Improved tools for biological sequence comparison; *Proc. Natl. Acad. Sci. USA* **85** 2444–2448

Rost B, Liu J, Nair R, Wrzeszczynski K O and Ofran Y 2003 Automatic prediction of protein function; *Cell. Mol. Life Sci.* **60** 2637–2650

Rychlewski L, Zhang B and Godzik A 1998 Fold and function predictions for *Mycoplasma genitalium* proteins; *Fold Des.* **3** 229–238

Schaffer A A, Wolf Y I, Ponting C P, Koonin E V, Aravind L and Altschul S F 1999 IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices; *Bioinformatics* **12** 1000–1011

Schroeder B G, Peterson L M and Fleischmann R D 2002 Improved quantitation and reproducibility in *Mycobacterium tuberculosis* DNA microarrays; *J. Mol. MicroBiol. Biotechnol.* **4** 123–126

Snider D E Jr, Raviglione M and Kochi A 1994 Global Burden of Tuberculosis; in *Tuberculosis: Pathogenesis, protection, and control* (ed.) B R Bloom (Washington DC: Am. Soc. Microbiol.) pp 3–11

Sonnhammer E L L, Eddy S R and Durbin R 1997 Pfam: A Comprehensive Database of Protein Families Based on Seed Alignments; *Proteins* **28** 405–420

Sonnhammer E L L, Von Heijne G and Krogh A 1998 A hidden Markov model for predicting transmembrane helices in protein sequences; in *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, California (eds) J Glasgow, T Littlejohn, F Major, R Lathrop, D Sankoff and C Sensen, pp 175–182

Sousa M C and McKay D B 2001 Structure of the universal stress protein of *Haemophilus influenzae*; *Structure (Camb)* **9** 1135–1141

Strong M, Mallick P, Pellegrini M, Thompson M J and Eisenberg D 2003 Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization a combined computational approach; *Genome Biol.* **4** R59 Epub

Tatusov R L, Galperin M Y, Natale D A and Koonin E V 2000 The COG database: a tool for genome-scale analysis of protein functions and evolution; *Nucleic Acids Res*. **28** 33–36

Thornton J M 2001 From genome to function; *Science* **292** 2095–2097

Voladri R K R, Lakey D L, Hennigan S H, Menzies B E, Edwards K M and Kernodle D S 1998 Recombinant Expression and Characterization of the Major *b*-Lactamase of *Mycobacterium tuberculosis*; *Antimicrob. Agents Chemother*. **42** 1375–1381

Wagner J, Lerner R A and Barbas C F 3rd 1995 Efficient aldolase catalytic antibodies that use the enamine mechanism of natural enzymes; *Science* **270** 1797–1800

Wootton J C and Federhen S 1993 Statistics of local complexity in amino acid sequences and sequence databases; *Comput. Chem.* **17** 149–163

Zhu H and Riggs A F 1992 Yeast Flavohemoglobin is an Ancient Protein Related to Globins and a Reductase Family; *Proc. Natl. Acad. Sci. USA* **89** 5015–5019

MS received 13 February 2004; accepted 15 June 2004

Corresponding editor: VIDYANAND NANJUNDIAH