Structural Annotation of the \textit{Mycobacterium tuberculosis} Proteome

NAGASUMA CHANDRA, SANKARAN SANDHYA, and PRAVEEN ANAND
Department of Biochemistry, Indian Institute of Science, Bangalore, Karnataka 560012, India

\textbf{ABSTRACT} Efforts from the TB Structural Genomics Consortium together with those of tuberculosis structural biologists worldwide have led to the determination of about 350 structures, making up nearly a tenth of the pathogen's proteome. Given that knowledge of protein structures is essential to obtaining a high-resolution understanding of the underlying biology, it is desirable to have a structural view of the entire proteome. Indeed, structure prediction methods have advanced sufficiently to allow structural models of many more proteins to be built based on homology modeling and fold recognition strategies. By means of these approaches, structural models for about 2,877 proteins, making up nearly 70\% of the \textit{Mycobacterium tuberculosis} proteome, are available. Knowledge from bioinformatics has made significant inroads into an improved annotation of the \textit{M. tuberculosis} genome and in the prediction of key protein players that interact in vital pathways, some of which are unique to the organism. Functional inferences have been made for a large number of proteins based on fold-function associations. More importantly, ligand-binding pockets of the proteins are identified and scanned against a large database, leading to binding site–based ligand associations and hence structure-based function annotation. Near proteome-wide structural models provide a global perspective of the fold distribution in the genome. New insights about the folds that predominate in the genome, as well as the fold combinations that make up multidomain proteins, are also obtained. This chapter describes the structural proteome, functional inferences drawn from it, and its applications in drug discovery.

Deciphering the complete genome sequence of \textit{Mycobacterium tuberculosis} in 1998 (1) marked an important milestone in tuberculosis (TB) research and has triggered a whole array of downstream research in the area. A bewildering array of omics data of the organism and genome sequences of related species is now available (2–5). One of the major obstacles to ready exploitation of such huge volumes of genomics data is the lack of annotation of many of the gene products. With the architectural blueprint of the organism at hand, a logical next step is to comprehend the huge pool of data to identify and understand the function of the individual gene products.

From a protein function standpoint, the only practical way to convert vast quantities of raw sequence data into meaningful information is through transfer of annotation from known related proteins. Indeed, advances in bioinformatics approaches have become more reliable in transferring functional annotation and integrating sequence and protein family classifications (6, 7). However, a finer appreciation of the molecular mechanisms within the cell is possible only through the elucidation of protein structure. It is clear that three-dimensional protein structures, where available, provide much better functional insights than one-dimensional sequences. Protein structures also pave the way for the use of much more sensitive approaches for detecting similarities.

\textbf{Received:} 6 August 2013, \textbf{Accepted:} 3 October 2013, \textbf{Published:} 28 March 2014

\textbf{Editors:} Graham F. Hatfull, University of Pittsburgh, Pittsburgh, PA, and William R. Jacobs, Jr., Howard Hughes Medical Institute, Albert Einstein College of Medicine, Bronx, NY


\textbf{Correspondence:} Nagasuma Chandra, nchandra@biochem.iisc.ernet.in

\(©\) 2014 American Society for Microbiology. All rights reserved.
among proteins (8, 9). Even for those sequences where the biochemical function is known, knowledge of the protein three-dimensional structure can provide further insight into the details of function and mechanism. The structure can reveal unique features of subunit and quaternary associations, substrate recognition, allosteric regulation, and the identification of putative active site residues. These aspects of structural information can be readily applied during the process of drug design (10).

Hence, in combination with sequence-based genome annotation efforts, structural genomics projects have launched large-scale studies to determine unique structures of proteins in an organism, primarily through X-ray crystallography (11). Toward the goal of obtaining structural information on a large number of proteins within the genome, a specific structural genomics consortium was formed for TB about a decade ago, the Tuberculosis Structural Genomics Consortium (TBSGC) (12). Efforts from such a consortium of several investigators across the globe, along with other TB structural projects, have led to the elucidation of structural information for about 10% of the proteome. The TBSGC has contributed 724 structures from 266 unique open reading frames to date, accounting for one-third of M. tuberculosis protein structures in the Protein Data Bank (PDB) (13). Although tremendous advances have been made in experimental methods to determine protein structure on a high-throughput scale, some bottlenecks still persist. These include experimental limitations of determining the right conditions for proteins that do not crystallize easily or occur in multidomain contexts or as large protein assemblies. Membrane proteins are also difficult to crystallize, and consequently, these efforts have been more successful for single-domain, cytoplasmic proteins (14).

On the other hand, the availability of a large number of structures of proteins from several organisms in the PDB (15), together with the advances in computational structure prediction algorithms, makes it feasible to obtain structural models for many more M. tuberculosis sequences with reasonable confidence (16). This approach has been utilized to obtain structural information for nearly 70% of the proteome (17). There are already a few excellent review articles on the structural biology of M. tuberculosis describing structural details of proteins involved in important biological processes (12, 13, 18, 19). This chapter focuses on a structural view of the entire proteome and the functional inferences obtained through this knowledge. Structure-based function annotation is discussed along with examples of insights into the mechanisms by which various biological events take place. The annotation includes identification of conserved residues in the family of proteins, identification of various sequence and substructural motifs, functional associations, compatible quaternary structures, and possible ligands and is expected to serve as a useful resource for TB researchers.

**OBTAINING A STRUCTURAL PROTEOME**

Sequence-based methods are used quite extensively for identifying homologous proteins (20). Homology-based comparative modeling has become a standard method to obtain the three-dimensional structural details of the protein (16). Broadly, homology models of the individual proteins are built based on rigorous, sometimes iterative steps of template identification for selecting the best template and alignment verification, followed by subsequent steps of structure validation. Methods in structure prediction have over the last two decades improved tremendously in applicability, speed, and confidence, as judged by the community-wide CASP experiments (21).

Where the extent of similarity is high, most methods usually detect homologs well. In cases of lower similarity, detection becomes a bit more complex and has to rely on sensitive approaches. Methods that use sequence-structure relationships can be broadly classified into two categories: profile-based and threading approaches. Profile-based approaches capture the residue environments of the proteins belonging to a particular superfAMILY and encode them in position-specific scoring matrices (22). Traditional sequence-based approaches are then adopted using position-specific scoring matrices to detect the homologs. In some cases where the homology is not readily detectable, HMMs (hidden Markov models) are used to find remote homologs (23). Threading, alternatively, fits the query sequence onto a database of known protein folds and evaluates the compatibility of the sequence with a structure using empirical scoring functions derived from a database of high-quality protein structures (24). Once the homology with a known structure is detected, the complete three-dimensional structure of the protein can be determined through homology modeling, a computational approach to determine the three-dimensional structure from the alignment of a query protein with a similar template protein whose structure has been solved. Whether homologous proteins are likely to perform similar functions can be further verified by examining conservation of residues that are involved in reaction chemistry.

Once the three-dimensional structure of the protein is predicted, rigorous evaluation must be carried out to
check the quality of the models. The verification criteria generally used are (a) statistical potential scores (25) of the model, which reflect compatible residue environments of the query sequence and structural compactness; (b) statistical significance of alignments and extent of sequence similarity obtained from routinely employed sequence alignment tools such as BLAST (26) and PSI-BLAST (27); (c) deviation from ideality in geometric parameters (28), which considers bond lengths, bond angles, planarity, and dihedral angles; (d) deviation in the protein backbone dihedral angles \( \varphi, \psi \) from allowed regions in the Ramachandran plot; (e) statistics of nonbonded interactions between different atom types (29); and (f) conservation of residue contacts, solvent accessibility profiles, and secondary structure compatibility (30).

Using the above-mentioned computational protocols, it is now feasible to obtain structural information at a genome scale. Automatically generated models using the Modeller package are made available in a dedicated database called MODBASE (31). For around 70% of the M. tuberculosis genome, structural models with reasonable confidence have been generated (17) (Fig. 1). This information has enriched our understanding of the pathogen by providing insights in atomistic detail about various pathways and biological processes, as described below.

**FUNCTIONAL INSIGHTS FROM PROTEIN STRUCTURES**

Obtaining a protein’s structure is the first step in the journey of obtaining mechanistic insights into the function of that protein. Experimental characterization of a protein function is available only for some proteins. However, the structural data resource is growing rapidly. The need to navigate and comprehend this large resource of both experimental and theoretical structural data has led to the emergence of a discipline called structural bioinformatics (11, 32, 33), which has truly come of age in the last decade (34). Structural bioinformatics is probably best thought of as an approach that rationalizes and classifies information contained in the three-dimensional structures of molecules, in terms of their functional capabilities, thus ultimately helping in understanding how they function in atomic detail. A main advantage of this approach over simpler sequence-based methods is that besides helping associate a molecule with a function, it also provides ultimate insights into the mechanisms by which various biological events take place. Thus, structural bioinformatics can serve as a bridge in transforming protein structures into biological insights.

An annotation exercise typically involves tasks of recognizing similarities and relationships among proteins, deriving structural patterns, correlating with function, and ultimately utilizing such patterns for prediction. This can be achieved at different levels. First, the protein structure itself can be classified and analyzed under different levels of structural organization, which are typically provided by classification schema as in the SCOP (Structural Classification of Proteins) (35) and CATH (36) databases. The SCOP database, a manually curated database, stores the structural information in four levels of hierarchy, namely, (a) class: composition of secondary structures in proteins, (b) fold: a specific arrangement of secondary structural elements, (c) superfamly: domains of the fold that share similarities in function but poor sequence homology, and (d) family: domains in a superfamily that have recognizable sequence similarity and clear ancestry. Similarly, the CATH database classifies protein domains into class, architecture, topology, and homologous superfamly in a semiautomated way. Once a new structural model is placed in one of the known structural classes and families, it can be explored if functional information from other well-characterized members in that family can be transferred to the new protein.

It must be noted that the term “function” itself can be defined at different levels of protein structure hierarchy, as mentioned above. The function of a protein could in order of ascending hierarchy be (a) the molecular function and (b) the biological process it is involved in. For example, the function of the RecA protein could be described as ATP-binding and DNA-binding at the first level and as a component of homologous recombination and DNA repair at the second level. “Fold to function” models, which have been the basis for the functional annotation of proteins in some cases, typically relate a structural fold to a predominant function known to be performed by proteins containing that fold. On the one hand, high structural similarity between two proteins along their entire polypeptide chains generally indicates high similarity in function at both the molecular function and biological process levels. On the other hand, when part similarity is exhibited by two proteins in their structures, their functions need not necessarily be the same. Detailed studies would be required to infer function in such cases.

Broadly, part similarity can mean that either medium-to-high similarity exists only in a portion of the polypeptide chain, indicating the presence of a common domain, or that low-to-medium level similarity exists in most of the polypeptide chain. In many cases, in the first
FIGURE 1 Schematic diagram of *M. tuberculosis* proteome structural annotation. The figure depicts the circular map of the *M. tuberculosis* H37Rv structural proteome, corresponding to the first genome view reported with its complete genome sequence (1). The outer circle represents the model coverage in terms of the percentage of the polypeptide chain, whereas the inner circle represents the percentage sequence identity shared by each model with its corresponding template. On both the outer and inner circles, radiating lines are drawn to indicate the parameters of the structural model for the corresponding protein in the genome view. The length of the lines in both cases is proportional to their values in percentages. The 100% mark is also shown for both the circles. In the outer circle, those models that had greater than 40% length coverage are drawn outside the circle, whereas those with coverage of less than that are drawn inside the circle (for clarity). Length coverage is divided into five classes and color coded as indicated, while the levels of sequence identity are divided into four classes and color coded as indicated. Predominantly occurring folds in the proteome are shown surrounding the outer circle, ordered clockwise by frequency of occurrence (indicated in parentheses). doi:10.1128/microbiolspec.MGM2-0027-2013.f1
category, inferring the molecular level function for that domain alone would be possible, but inferring the biological process level function would be much harder. For the second category, however, functional inference at either level would not be meaningful. This is because fold-level similarity does not necessarily indicate similarity in the functional regions of the molecule.

Gene ontology (37) can be very useful when describing the putative function of a protein through biologically relevant key terms. Various studies have been carried out to annotate proteins in terms of functional key words at different levels of structural hierarchy (38, 39). The functional terms are also obtained from other sequence databases such as InterPro, Pfam, and Swissprot. These associations help us analyze fold-based function annotation at a genome scale. In M. tuberculosis the majority of superfamilies (44%) are associated with metabolism, followed by superfamilies involved in general multiple functions (14%) such as interactions with proteins, ions, lipids, and small molecules; the third highest is superfamilies involved in intracellular processes (10%) such as cell motility, cell division, cell death, intracellular transport, and secretion.

Structural information has been useful in gaining a functional understanding of proteins in many cases (40, 41). The PE and PPE families—named after the identification of PE (proline-glutamate) and PPE (proline-proline-glutamate) motifs present at the N-terminal end of proteins—are associated with antigenic variation in mycobacteria. Based on a computational genomics prediction that some members of the PE and PPE families are functionally linked, experiments designed to test their interactions confirmed their interaction in vitro by co-expression and copurification, following which the crystal structures of the PE25/PPE41 complex were successfully determined. Furthermore, based on the structural features of the complex, the pair of proteins was suggested to represent a docking site for an as yet unidentified bacterial or host target and likely to be involved in signal transduction (42). Other important examples include (a) the identification of Rv3547 as a nitroreductase (43), (b) annotation of Rv2175c as a DNA-binding protein and a substrate of a protein kinase PknL, (c) understanding the subcellular localization of Rv2626c, a hypoxic response protein, and its involvement in modulating macrophage effector functions, (d) identification of Rv1155 as an enzyme in the pyridoxal-5′-phosphate biosynthesis pathway (44), (e) identification of a possible common mechanism of action for cyclopropane syntheses (PcaA) (45) and methyl transferases (CmaA1 and CmaA2) (46), and (f) detection of Rv1347c as an enzyme in the mycobactin biosynthetic pathway, thus filling the missing link in the pathway (47). One of the most interesting insights concerns Rv3361c, MfpA, whose structure revealed a fold labeled as a right-handed quadrilateral beta-helix, which resembles standard B-DNA in size, shape, and electrostatics (48). Structural studies explained how the protein binds to DNA gyrase, mimicking DNA and affecting binding with fluoroquinolones, thereby conferring drug resistance (48).

Functional annotation from structures in a more automated manner can be quite complicated, since one-to-many and many-to-one relationships among fold types and functional categories are known to exist in a number of cases (49, 50). A more direct and insightful method is to extract functionally important regions in proteins and associate these regions with particular function(s) through substructure comparisons. The need to understand functional sites is not restricted to new sites. Even where protein structures are determined crystallographically as a complex with a ligand, a complete description of their binding sites is not always obtained, because they may not be cocrystallized with all the ligands required for the function of the molecule or because the bound ligands are often substitutes of the natural ligands. Identification of all relevant binding sites in protein molecules, therefore, forms a key step toward gaining functional insights from protein structures.

The details of functional inference that can be derived from protein structures through computational annotation are directly dependent upon the quality of the structure. Protein structures for about 1,064 proteins in the M. tuberculosis genome could be modeled based on a template with a sequence identity of ≥30%. For such structures, more detailed functional analysis can be performed to predict ligand-binding sites, active sites, conserved surfaces, ligand associations, and nucleic acid interactions. Since no single method can accurately predict the function of the protein from its given structure, a function annotation pipeline involving more than one algorithm is more useful. An example functional annotation for Rv1485 (hemZ) is depicted in Fig. 2.

**COVERAGE OF M. TUBERCULOSIS STRUCTURAL PROTEOME AND ANNOTATION**

The coverage of the structural information currently available for M. tuberculosis extends to about 70% of the proteome. Analysis of the distribution of structural information into various TubercuList (5) functional
categories reveals that the topmost functional category with maximum structural information coverage is the information pathway (95.23%), followed by intermediary metabolism and respiration (94.76%) and lipid metabolism (91%). Interestingly, only 219 unique folds were associated with all the proteins involved in metabolism. Even in the category of conserved hypotheticals, models could be built for about 50% of them. The functional categories for which the structural information is limited are insertion elements and phages, due to lack of homology with any experimentally derived protein structures in the PDB. Around half of the genes involved in cell wall processes also do not have structural information because they would fall into the category of membrane proteins, which are inherently tough to crystallize and obtain structural information from.

Overall, there exists structural information for about 1,097 enzymes, and for 647 of them the active site residues could be determined through structure-based function annotation. Structure-based function annotation also revealed 1,728 ligand associations for the protein structures through binding site analysis and comparison. Around 740 possible DNA-interacting proteins were found to be present in the structural proteome.
FEATURES OF THE M. TUBERCULOSIS STRUCTURAL PROTEOME

The availability of the protein structures at a genome scale of M. tuberculosis presents a unique opportunity to understand the fold space covered by this proteome and to understand if there are particular biases or preferences. Figure 3 gives the distribution of various families of the proteins that have been modeled and the coverage of models depending upon the Tuberculist classification.

Fold Distribution

The availability of the structural models of a particular organism, on a large scale, can give us the opportunity to analyze the distribution of structural classes and the most commonly occurring folds. A comparison with the currently existing SCOP database reveals that all the seven major structural classes of SCOP are observed among the available models of the M. tuberculosis proteome as well, although a marginal preponderance to alpha/beta class (36.45% after normalizing it to number of proteins when compared to 14.45% in SCOP) can be seen in M. tuberculosis. Out of the 1,195 known folds present in the SCOP database, 351 are seen to be present in the M. tuberculosis proteome, amounting to a coverage of known folds to about 30%. Similarly, at the family and superfamily levels, a significant coverage of 20% and 24.36% was observed in the M. tuberculosis proteome.

Around one-third (887 proteins) of the structural proteome is dominated by the most commonly occurring folds such as (Fig. 1) DNA/RNA-binding 3-helical bundle, P-loop containing nucleotide triphosphate hydrolases, TIM beta/alpha barrel, NAD(P)-binding Rossmann fold, SAM-dependent methyltransferases, ferredoxin-like, and alpha/beta-hydrolases. This is expected because it is known that distribution of copies of folds per genome in bacteria is known to follow power law, and these folds are commonly observed in most bacteria. But in the M. tuberculosis modeled proteome, ferritin fold, adopted by the N-termini of the PE and PPE proteins, dominates the list, which is not surprising given that nearly 180 of the proteins of this family exist in M. tuberculosis. Members of the PE and PPE families are known to be cell surface molecules involved in regulation of dendritic cell and macrophage immune-effector functions and also as antigens generating strong humoral responses that are possibly important for antigenic variability.

Fold Combination

The modeled structural proteome could also give us insights about combinations of protein folds that come together to form a polypeptide chain. Around 489 multidomain proteins could be identified in the modeled proteome. The fold combinations utilized can be represented in the form of a network with nodes representing a fold and edges representing the co-occurrence of folds within a polypeptide (Fig. 4A). The network revealed that the P-loop containing nucleotide triphosphate hydrolases is the fold with the highest number of co-occurrences. The most commonly occurring fold combination is observed between a tetra-cyline repressor-like fold and a DNA/RNA-binding 3-helical bundle. This combination is usually observed in the proteins that are known to play an important role in ribosomal protection and to help in the regulation of various efflux proteins.

LEVELS OF ANNOTATION

The availability of protein structures for mycobacterial proteins provides an ideal starting point for arriving at functional insights at different levels. With improvements in bioinformatics approaches, many databases available online provide automated functional annotation of proteins. The availability of structures can complement such efforts to understand the molecular basis of biological function. Automated pipelines that impose strict quality filters to model the M. tuberculosis proteome, for instance, have resulted in structure assignment for 2,877 proteins. Using a case study, we describe how such pipelines can be applied to associate a structure with a protein and improve current understanding of its function.

Function Annotation through the Structural Annotation Pipeline

Rv1485 (P0A576: 344 residues, 4.99.1.1), annotated as a ferrochelatase, is known to catalyze ferrous insertion into Protoporphyrin IX to form Proto-heme. Based on sequence homology, Rv1485 was modeled using 1HRK (A chain, human ferrochelatase, 359 residues) as the template. The derived model, when superimposed on the template, showed less than 0.9 Å root mean square deviation, suggesting that the input alignment was of good quality. Parameters such as extent of residue conservation, compliance with the Ramachandran plot, and ERRAT scores could be determined to assess the quality of the generated models. For this example, these were all found to be satisfactory. Subsequently, one could perform a multiple-sequence alignment by including homologs from related bacteria, yeast, and humans, and such alignments showed a high conservation of residues in the protein core, indicating a well-conserved
FIGURE 3 Coverage of a structural proteome. (A) Distribution of different structural classes as described through folds, superfamilies, and families occurring in SCOP and correspondingly in the *M. tuberculosis* proteome adjacent to it. (B) Distribution of structural information according to TubercuList functional categories. The inner circle represents the total number of genes in a particular functional category, and the outer circle represents the genes with structural information in the corresponding functional category. doi:10.1128/microbiolspec.MGM2-0027-2013.f3
fold (Fig. 2C). Eukaryotic ferrochelatases typically possess three regions: an N-terminal organelle targeting region that is proteolytically cleaved, a second core region of 330 residues sharing homology with bacterial ferrochelatases, and a C-terminal region that contains the dimerization motif as three of the four cysteine ligands of the 2Fe-2S cluster. These enzymes are known to function by involving the 2Fe-2S clusters. The availability of the alignments is then useful to compare the cysteine ligands of the 2Fe-2S clusters across the diverse species. In *S. pombe* ferrochelatase, for instance, cysteines analogous to the eukaryotic ferrochelatases are found. *M. tuberculosis* counterparts are seen to possess four cysteine-ligating residues, at C158, C332, C339, and C341.

Detailed structural analysis of the template has shown that it exists in an open substrate-free and closed substrate-bound form. In the two forms, it was shown that the active site pocket is closed around the porphyrin macrocycle. A number of active site residues are known to undergo reorientation of side chains due to a hydrogen bond network involving H263, H341, and E343. PocketDepth (56) and LigsiteCSC (57) predictions made

---

**FIGURE 4** Fold combination and higher-order assemblies. (A) Network of fold combination observed in modeled multidomain proteins. Each node represents a fold, and an edge represents two folds occurring together within a polypeptide. The topmost occurring fold combination is the tetracycline-like repressor C-terminal (a.121) domain with DNA 3-helical bundle (a.4). An example protein, Rv3557c, is shown with both folds highlighted. (B) Examples of higher-order assemblies. The predicted assembly of methylmalonyl CoA mutase derived from the structural template 1REQ is shown. The assembly consists of Rv1492, MutA (cyan), and Rv1493, MutB (green). The conserved residues that could be involved in the interaction at the interface are shown in stick representation below. Similarly, the complex of fumarate reductase generated from 1KF6 is shown below, with residues involved in quinol binding that are conserved highlighted as spheres. doi:10.1128/microbiolspec.MGM2-0027-2013.f4
on the modeled protein identified pockets that overlapped with the template pockets harboring the 2Fe-2S cluster and the cocrystallized ligand (cholic acid, 1HRK:A) (Fig. 2D). Comparison of the pockets in the modeled structure with that of the template also showed a similar network of residues in mycobacterial ferrochelatases. Methods such as ProFunc (58), when applied independently, also predicted these pockets. Finally, an overlap of the binding sites through methods such as PocketAlign (59) (Fig. 2F) could then be used to demonstrate the extent of structural overlap in the binding sites. In this example, for which function as a ferrochelatase was predicted, a combination of the results from multiple methods on the modeled structure could complement and enrich information pertaining to their binding pockets and reaffirm their role as potential ferrochelatases.

The level of detailed annotation possible for each protein, given a model, varies depending on the extent of homology to characterized proteins, homologous structures, and associated binding sites. In some cases, available information of the protein may be enriched further, while in others finer details of annotation such as identifying substrate specificities, etc., may be possible, especially where a template selected with high confidence and homology harbors a potential ligand. More importantly, the availability of models has permitted annotations of function for completely uncharacterized proteins in a few cases. In at least 39 cases, higher-order interactions and assemblies could also be inferred by examining gene neighborhoods of the target proteins. In the following section, different levels of function annotation are highlighted using representative examples.

**Prediction of higher-order assemblies**

Rv1552, a potential fumarate reductase, was modeled with our approach, using 1KF6:A as a template. The model generated for this protein covered 98.6% of the protein and showed 55% sequence identity with respect to 1KF6:A. Furthermore, 1KF6 was found to be one of the four subunits of the quinol-fumarate reductase respiratory complex of *Escherichia coli*, which is an integral-membrane complex. Interestingly, structural annotation of the gene neighbors of Rv1552, namely Rv1553, Rv1554, and Rv1555, identified 1KF6:B, 1KF6:C, and 1KF6:D as templates, suggesting that a similar four-subunit assembly is likely in *M. tuberculosis* as well.

It is known that in *E. coli*, 1KF6 catalyzes the final step of anaerobic respiration when fumarate is the terminal electron acceptor (60). During this type of anaerobic respiration, electrons are donated to quinol-fumarate reductase by menaquinol (MQH2) molecules in the membrane. The electrons are transferred to a covalently bound flavin adenine nucleotide at the active site through three distinct iron-sulfur clusters and ultimately are used to reduce fumarate to succinate. Two of these chains, the flavoprotein (FrdA) and the iron protein (FrdB), compromise the soluble domain, which is involved in fumarate reduction, whereas the remaining two subunits (FrdC and FrdD) are membrane-spanning polypeptides involved in the electron transfer with quinones.

Results of the SCOP domain assignments of the *M. tuberculosis* proteins Rv1552, Rv1553, Rv1554, and Rv1555 show domain assignments and combinations that are observed in each of the template subunits with high statistical confidence (<10 to 20). Furthermore, alignments of the individual *M. tuberculosis* proteins with their closely related structural templates also show high similarity between these two proteins. Two quinol-binding sites in the membrane-spanning region of the template designated as Qp and Qd are known to lie proximal and distal to the 3Fe:4S cluster. Key residues involved in the quinol-binding site of the complex include C210, Q225, and K228 (B chain); F24, R28, and E29 (C chain); and R81, H84, H80, F17, E10, and W14 (D chain). These are also observed in the four-subunit assembly proposed by our pipeline (Fig. 4B), suggesting similar function and assembly in the *M. tuberculosis* proteins.

In another example, Rv1493 (P65487), currently annotated as a probable methylmalonyl-coenzyme A (CoA) mutase large subunit, was modeled using 1REQ:A as the template (71% identity and 98.27% model coverage). Methylmalonyl-CoA mutase catalyzes the isomerization of succinyl-CoA to methylmalonyl-CoA during synthesis of propionate using adenosylcobalamin as the cofactor (Fig. 4B). Alignments of the modeled protein with the template show conservation of active site residues such as Y89, H244, K604, D608, and H610 in the model as well. PocketDepth and SURFnet predictions of a highly conserved active site and a cobalamin-binding site in the model are seen in the template at equivalent positions. Gene neighborhood studies show that Rv1492, modeled with 1REQ:B as the template, lies upstream of Rv1493. With the availability of models from both these proteins, studies can be undertaken to predict the interacting residues lying at the interface of these two interacting chains. Detailed analysis of such interfaces may point to useful leads in the
design of molecules that target such interfaces to disrupt protein function.

**Improved annotation of conserved hypothetical protein**

Rv3402c is a conserved hypothetical protein with an unknown function and structure. Our annotation pipeline identifies an aminotransferase, 1MDO, that catalyzes the transfer of an amino group from glutamic acid to a UDP-linked ketopyranose molecule as the template. This aminotransferase, a homodimer, is involved in a lipid A modification pathway that in some bacteria, such as *Salmonella*, confers resistance to cationic antimicrobial peptides. Crystal structures of ArnB show the

**FIGURE 5** Example annotations. (A) Example annotation of a conserved hypothetical protein, Rv3402c. The superposition consists of a model of Rv3402c (shown in green), and the template 1MDO is shown as cyan. The conserved residues predicted to be involved in the interaction with PLP are represented as sticks. (B) The structure-based sequence alignment of Rv3402c with the template 1MDO. Functionally important residues are marked with (*). The amino acids are colored based on their chemical properties. (C) Superposition of Rv0469 (green) with the template 1KPG. The residues involved in cofactor recognition are shown in blue, and residues determining the substrate specificities are highlighted in red. The predicted pocket is shown in a surface representation. (D) Superposition of Rv2503c (green) with the template 2CTZ. The pocket predicted using PocketDepth and SURFnet enclosing the active site is shown in a mesh. doi:10.1128/ASMscience.org/MicrobiolSpectrum.MGM2-0027-2013.f5
presence of a large N-terminal cofactor-binding domain and a smaller three-stranded beta sheet. The template structure in the presence of cofactor, product, and inhibitor has revealed key residues that are implicated in the aminotransferase reaction. Comparison of the sequence of Rv3402c and the template shows high similarity, as does overall structural superposition (root mean square deviation of 0.49 Å). Residues that interact with the cofactor pyridoxal phosphate (PRP) in the template include Y136, H329, H163, K188, S183, H185, E194, T64, W89, D160, and F330. These residues are seen in structurally equivalent positions in Rv3402c. Likewise, residues that interact with the substrate 1-glutamate in ArnB are also seen in Rv3402c, suggesting likely aspartate aminotransferase activity in M. tuberculosis. Figures 5A and B show an alignment of Rv3402c and 1MDO:A from the pipeline. Residues that are involved in cofactor binding in 1MDO:A and in the model are highlighted.

Enrichment of functional annotation
Rv0469, a possible mycolic acid synthase, has been modeled in Modbase with 1KPG:A, a cyclopropane-fatty-acyl-phospholipid synthase 1, as the template. Mycolic acids are major components of the cell wall of M. tuberculosis, and several efforts are under way to examine their potential as a drug target. Functional groups in the acyl chain of mycolic acids are understood to be important for bacterial pathogenesis and persistence. It is now known that there are at least three mycolic acid cyclopropane synthases (PcaA, CmaA1, and CmaA2) that produce these site-specific modifications. As in other methyltransferases, 1KPG:A adopts the seven-stranded alpha/beta fold and possesses a binding site for its cofactor, S-adenosyl-l-methionine. Multiple sequence alignments of Rv0469 with other known mycolic acid cyclopropane synthases such as CmaA2 (P0A5P0), PcaA (Q7D9R5), MmA2 (Q79FX6), MmA3 (P72027), MmA4 (Q79FX8), and CmaA1 (1KPG:A; P0C5C2) show a high conservation of residues involved in binding S-adenosylmethionine (cofactor), as do residues involved in ligand interactions. Ligand-binding-site predictions using PocketDepth and cleft predictions of the ProFunc server, when applied to the modeled Rv0469, identify potential substrate-binding and cofactor-binding sites with high confidence. Indeed, 86% of the cofactor-binding residues in 1KPG:A are seen at topologically equivalent positions in Rv0469. E124 that shows hydrogen bonds with N6 of S-adenosyl-l-homocysteine (SAH) is replaced with D123 in Rv0469. Assessments of the substrate-binding pocket show that 85% of the ligand-binding residues in 1KPG:A are also conserved in Rv0469 (Fig. 5C). Furthermore, cation-π interactions mediated by Y33 to stabilize the carbocation intermediate in the template are conserved (Y32) in Rv0469 as well. Since the ligand-binding pocket is 9 Å from the surface, it has been suggested that longer chain length mycolic acids can also be accommodated into the pocket. Such structures are thus ideal starting points for more detailed assessments involving the actual substrate (mycolic acids of varying chain length) within the predicted ligand-binding pockets of Rv0469.

Augmenting confidence of sequence-based annotation and confirming existing annotation
Rv2503c (P63650) is currently annotated as a probable succinyl-CoA:3-ketoacid-CoA transferase β subunit in Tuberculist. This protein was modeled by considering 1OOY:A, with which it shows 58% identity, as the template. 1OOY also functions as a succinyl-CoA:3-ketoacid CoA transferase. Even though the template lacks a bound ligand, biochemical studies show that N281, Q99, G386, and A387 are conserved in CoA-transferases, and correspondingly, these residues were also conserved in the model of Rv2503c. Furthermore, binding site residues predicted by both PocketDepth and SURFnet show that these residues lie within the predicted binding pocket of Rv2503c (Fig. 5D).

STRUCTURAL PROTEOME AND ITS IMPACT ON PATHOGENESIS AND VIRULENCE
Interactions between proteins are important for a majority of biological functions. For instance, signals from the exterior of a cell are transmitted to the cell interior through interactions between proteins in processes such as signal transduction. Likewise, the process of transcription, its regulation, and a number of other processes in the cell are a consequence of protein-protein interactions that are either obligatory or transient in nature. Capturing such interactions is key to understanding which proteins interact with each other. A number of experimental procedures such as yeast two-hybrid assays and coaffinity purification have been developed to infer linkages between proteins on a genome-wide scale. Computational methods that complement these techniques include the Rosetta stone, phylogenetic profile, and operon or gene cluster methods. Protein linkages identified by these methods reveal proteins that participate in
protein complexes, and therein further inferences about key interaction points in pathway(s) and related functions may be made. Therefore, such methods have been very useful in critically narrowing down the protein players involved in such interactions. Finer details on which specific parts interact and what kinds of chemical bonds mediate that interaction can be gleaned through the availability of protein structures. Structures of such proteins and their complexes can assist in the design of molecules that target such interfaces and disrupt their formation, as has been shown in the case of class III adenyl cyclases and other molecules involved in host-pathogen interactions such as NarL, Rv2413c, and LprG (70).

INSIGHTS INTO TWO-COMPONENT SYSTEMS

A key feature of all living cells is the ability to sense environmental signals and implement adaptive changes. In pathogenic bacteria, this ability to adapt and respond to external stimuli, through simple phosphorylation events between membrane-localized kinases and response regulators that activate or repress transcription, cascades a series of signaling events in the pathogen referred to as two-component systems (71). These defense systems are believed to play a vital role in early intracellular survival of the pathogen as well as in various aspects of virulence. Mutation studies show that in vivo survival of the pathogen in the host macrophages is affected for mutants of these proteins. With the process of phosphotransfer from histidine to aspartate in such systems being absent in humans, such two-component systems are being increasingly viewed with interest in the area of structure-based drug design (72). Due to the nature of the interactions mediated by these proteins, such proteins are necessarily multidomain in nature, with long flexible linkers connecting the domain, and therefore, crystal structures of such proteins have always been challenging. Small-angle X-ray scattering approaches have been applied quite successfully to determine many of these structures. In mycobacteria, 13 structures are known to date in the PDB; these include MtrA (2GWR), NarL (3EUL), PdtaR (1SN8), PrrA (1YS6, 1YS7), and RegX3 (1ZKV).

The use of a number of bioinformatics approaches to predict the number of such components in mycobacteria has shown that they do not have a very high representation of such systems (71). Of the total 55 proteins in M. tuberculosis that are implicated in these systems in KEGG, 52 have been structurally modeled. Of the 12 complete two-component systems that are known to be encoded in the mycobacterial genome, 8 complete systems involving the sensor and the regulator have been modeled through such pipelines. This includes four of the five two-component systems, namely, SensX3-RegX3, PrrA-PrrB, MprA-MprB, MtrA-MtrB, and PdtaR-PdtaS, conserved in all mycobacterial species (19) sequenced to date. Furthermore, three of the five two-component systems, conserved in all except Mycobacterium leprae, namely PhoP-PhoR, KdpD-KdpE, TrcR-TrcS, DevR-DevS, and TrcX-TrcY, NarL have also been modeled through this approach. The availability of models for such complexes provides a useful starting point to determine the basis of such key molecular events and their biological roles in signal transduction systems and in adapting to host responses in pathogenesis.

Toxin-antitoxin (TA) gene pairs specify proteins that encode a stable toxin and an unstable antitoxin. Under normal physiological conditions both proteins are expressed and form a tight complex. However, under conditions of stress such as viral invasion, oxidative stress, or antibiotic attack, the unstable antitoxin is rapidly degraded by proteases, and thus the toxin is free to kill the host cell (13). The number of such TA systems in mycobacteria is high, and estimates range from 90 to over 120 genes. It is suggested that these protein pairs may play an important role in the persistence of M. tuberculosis and contribute to its survival within the phagosomal subcellular compartment following to infection. Structural genomics efforts have already obtained the crystal structures for some TA pairs such as VapBC systems and RelBE systems, and in both these systems the toxins are RNases. Of the 122 TA genes extracted from the TADB database (73), 62 have been modeled (17), 38 of which were seen to adopt the PIN-domain-like fold, a characteristic of nuclease enzymes that cleave single-stranded RNA. Four others contain the lambda-repressor-like DNA-binding domain and YefM-like domain. The availability of structural information for such TA pairs should improve our understanding of how antitoxins inhibit their toxins and describe better the molecular basis of their mode of action.

APPLICATION IN DRUG DISCOVERY

The availability of the protein structure opens up avenues for structure-based drug discovery. The impact of structural-level understanding is felt at a number of steps in the discovery pipeline. To begin with, structures of the relevant protein molecules provide a much more detailed level of understanding of the underlying
biological processes. Drug target identification benefits immensely from structural data. Properties desired in a target molecule, such as druggability, can be studied with the help of structures. Druggability in this context is a term that describes whether the given protein is amenable to modulation with a small molecule, which translates to asking whether a defined small molecule binding site can be identified in the protein and hence if the target is chemically tractable. Specificity to the pathogen can also be studied by combining structural information and informatics. The most direct and widespread application of structural knowledge, however, is in lead identification and lead refinement or lead optimization. Structural knowledge has also been used for searching through potential ligands among a database of approved drugs available for different diseases, setting the stage for possible drug repurposing as well. Each of these is described in the following section.

CURRENTLY EXPLORED TARGETS FOR STRUCTURE-BASED ANTI-TB DRUG DISCOVERY

The availability of a large number of structures in various functional classes from *M. tuberculosis* has provided the platform for the selection of targets from different pathways and evaluation of their druggability. These include protein targets involved in cell-wall biosynthesis such as alanine racemase (74); arabinogalactan and LAM biosynthesis pathways (61); and mycolic acid pathway players such as polyketide synthase Pks (75, 76), acyl-AMP ligase (77, 78), FadD32 (78, 79), the AccD4-containing acyl-CoA carboxylase (80), and others. Pathway comparison methods have highlighted a number of enzymes and pathways that are unique to mycobacteria and absent in the host. These include enzymes involved in essential amino acid synthesis such as enzymes from the shikimate pathway or pathways involved in vitamin B₂ and B₃ synthesis. Other targets include enzymes involved in regulatory processes, DNA metabolism–uracil DNA glycosylase (81), and carbon assimilatory processes such as the glyoxylate shunt pathways (18). TA pairs and other molecules that are thought to be involved in bacterial persistence are also of interest as drug targets (82). The availability of crystal structures of a number of such complexes such as VapBC and RelBE complexes (83) has not only shed light on the structural basis of such interactions, but have contributed to a better understanding of the competition exhibited by antitoxins for the RNA-binding substrates of the toxin. Likewise, comparison of toxins has shown that they differ extensively in substrate recognition sites, suggesting that different TA systems may be triggered by different regulatory controls and conditions.

NEW TARGET IDENTIFICATION: STRUCTURAL ASSESSMENT OF DRUGGABILITY

An important criterion to consider during drug target identification is that the target is chemically tractable and should be sufficiently different from that of any host protein (84). Chemical tractability, or druggability, can be translated into a problem of identifying whether a protein has a binding site capable of recognizing a small molecule ligand with reasonable affinity and whether the site and hence the binding is specific enough, especially compared to host proteins. Considering specificity at an early step of target identification itself is likely to lead to significant reduction in failure rates due to cross-reactivity of the designed compound between host and pathogen. Similarity between proteins has been considered for some time now by comparison of their sequences, but where structural data for both proteins are available, similarity is better captured through structural comparisons. One of the ultimate requirements for determining the pharmacological profiles of drug molecules is which protein(s) in the given physiological environment can recognize the given drug molecule, a question that can be addressed with the help of structural comparisons of all possible binding sites of the pathogen’s proteome with the counterpart set in the host. The goal here is to be able to weed out target candidates that share very high similarity with binding sites from the human “pocketome,” since targeting these may lead to adverse drug reactions, due to inadvertent binding with human proteins. For this objective, it is not very intuitive to look at structural classes and overall properties such as the structural family or secondary structural types or a similar descriptor. Instead, the study of binding sites in proteins will provide information on which proteins could bind a given drug.

To render it meaningful, this type of analysis needs to be carried out at the proteome scale. About 229 and 3,515 experimentally derived structures are available in the PDB from *M. tuberculosis* and human, respectively. In addition, as described earlier, structural models for 70% of the *M. tuberculosis* proteome have been built. Similarily, homology-derived models are available in the ModBase database for 16,000 human proteins, making it feasible to carry out a proteome-scale structural assessment of targetability. A data set of about 3,500
binding pockets—comprising known ones from PDB structures combined with a large number of consensus binding sites predicted using three methods in putative drug targets in *M. tuberculosis*—were compared with about 70,000 similarly assimilated pockets from host proteins using algorithms (88) written specially for such a task (89). A mega task of about 245 million pairwise comparisons was performed on the IBM BlueGene supercomputer in a week. Of the 767 proteins tested in this manner, 145 were eliminated due to closely matching pockets in the human proteomes. It is possible that some of the eliminated proteins have sites that share part similarity only and hence could be targetable by using structural data to clearly identify common residues and different residues in each set, but this requires a close and more detailed analysis of all the pockets in the protein. As many as 622 proteins in *M. tuberculosis* that are flagged as essential for bacterial growth were found to contain proteins with binding pockets for small molecules that were unique to the pathogen. The screen thus identified unique druggable pockets in the pathogen and has yielded a shortlist of targets for designing antitubercular drugs.

**STRUCTURE-BASED LEAD DESIGN**

Knowledge of the structure of the target macromolecule and the functional regions in it provides a detailed definition of the binding site(s) that needs to be targeted with a drug molecule to achieve the desired manipulation. In the case of receptors involved in signaling cascades, the manipulation can be in the form of a small molecule that works as either an agonist or an antagonist, depending on both the disease and the receptor. For enzymes and other classes of targets, manipulation occurs predominantly through an inhibitor that prevents the natural substrate from binding to that protein. Lead design becomes a much more rational exercise with the knowledge of the target molecule. The approach has been successfully applied in the design of anti-influenza compounds (87, 88) and anti-HIV compounds targeting HIV protease (89, 90). The design can be performed through (a) virtual screening of large libraries of small molecules using docking methods or (b) fragment-based discovery combined with experimental structure determination and iteration of design or rational design of analogs of a known ligand with a specific aim to include or exclude interactions with particular residues in the binding site.

This has been shown in TB for the design of inhibitors against isocitrate lyase that play a key role in survival of the pathogen in the latent form during a chronic stage of infection (91, 92). The absence of the glyoxylate shunt pathway in mammals makes this an interesting target, and a number of interesting inhibitors—many of them mimics of metabolic intermediates in the pathway, such as succinate—have presented potential candidates showing inhibitory activity in the 0.10 μM range. Similarly, the structure of dehydroquinase from *M. tuberculosis* with bound ligand and inhibitors explained the molecular details of the reaction that could be used to design better inhibitors (93). Isoniazid, a frontline drug used in the treatment of TB, is known to be a prodrug, which is converted to the active species by catalase, which subsequently inhibits the function of InhA, a key enzyme in the mycolic acid biosynthetic pathway, by reacting nonenzymatically with NAD* and NADP* to form isonicotinoyl nucleotide adducts. In addition to this, isoniazid was shown through crystallographic studies to bind to dihydrofolate reductase (DHFR) (94), and the structure of isoniazid-NADP-DHFR was employed in structure-based inhibitor design to maximize the interactions of substituted isonicotinoyl derivatives in the two subsites characteristic of the DHFR binding site.

Similarly, DNA ligases, which participate in replication and repair mechanisms, are broadly classified into NAD*-dependent and ATP-dependent ligases. Due to their critical function in replication and repair, the crystal structure of the adenylation domain of the NAD*-dependent ligase of *M. tuberculosis*, LigA, was employed to infer a different relative juxtaposition of the subdomains of the enzyme when compared to other bacterial ligases. Furthermore, the syn-conformation of AMP observed in this crystal structure appeared to mimic the conformation of AMP that is expected in the third step of the classical ligase reaction, because it is not covalently bound to the active site lysine. The availability of the crystal structure led to *in silico* screening of inhibitors and identification of a novel inhibitor class based on glycosylureides. These inhibitors are capable of distinguishing between the NAD*- and ATP-dependent ligases, thus making them potential leads toward obtaining class-specific inhibitors of these enzymes (95). Similarly, ligand-based virtual screening efforts have also been shown with chorismate mutase (Rv1885c) to identify potential inhibitors (96).

Although a number of molecules can show inhibitory activity *in vitro*, transforming these activities to a nanomolar inhibitor range *in vivo* without impacting their ADME/Tox profile is often challenging. A number of interesting candidates have been explored to design
more effective molecules against InhA (97). Several other examples are featured in the recent literature, such as the optimization of inhibitors of protein kinases such as protein kinase G by AX20017 (98). Fragment-based design has also been used to guide derivatization of a lead series of β-lactamase inhibitors against M. tuberculosis β-lactamase (BlαC) (99) and has been successfully in the screening of fragment libraries against pantothenate synthetase to detect candidates that can be further modified to generate potent inhibitors (100).

Using a structural proteome-wide drug-target network of M. tuberculosis (TB-drugome), constructed by associating putative ligand-binding pockets in structural models of M. tuberculosis proteins with the known drug-binding sites, several new associations between approved drugs and proteins have been suggested (70), paving the way for possible drug repurposing. In this work, by combining molecular modeling, structural bioinformatics, and network approaches, a drug-target network was constructed, which the authors named the TB-drugome. Two proteins were connected if they exhibited similarity in their binding pockets. Drugs known for one protein were associated with another if the two proteins were connected with an edge in the network. Using such an approach, the anti-Parkinson’s drugs entacapone and tolcapone were identified as possible inhibitors of InhA, a well-known target in M. tuberculosis, and were subsequently validated in an in vitro study as well (97).

CONCLUSIONS

Currently, there are structural models for nearly 70% of the proteome of M. tuberculosis. The availability of structural data has been invaluable in advancing our understanding of the biology of M. tuberculosis. Functional annotations through structures have been made to a large number of proteins, and new annotations have been obtained for a number of hypothetical proteins. The advantages of structure-based annotation are twofold: (a) assigning a function based on the structural information, and that of the binding sites in it, makes the assignments more foolproof, and (b) they simultaneously provide a mechanistic basis for the assigned function by identification of the feasibility of binding, binding modes, and key residues involved in the process.

Getting a structural glimpse of a large portion of the proteome provides unique opportunities to explore the fold universe present in a single cell, fold frequencies, and preferences. For example, it is seen that the entire metabolism in the mycobacterial cell is achieved through a mere 219 folds. It is interesting to note that all seven major structural classes of proteins are represented in the M. tuberculosis proteome. A global perspective of folds that predominate in the genome, as well as the various fold combinations that make up multidomain proteins, is also obtained, throwing light on biologically significant protein assemblies.

Various stages of drug discovery benefit enormously from the availability of structural information of the M. tuberculosis proteome. Known or predicted targets can easily be characterized at a detailed level with the help of structural models. New target prediction is achieved not merely by considering the function of a molecule of interest but also by incorporating criteria of specificity and druggability, to pick targets that are unique to the pathogen and can be readily taken forward to the next steps in the pipeline. Through structural models, specificity is understood at a very high level of resolution by considering uniqueness in the binding sites compared to thousands of proteins in the host proteome, setting a new trend of increasing likelihood of drug safety at a very early stage in the drug discovery pipeline. Identification of about 10% of the proteome as putative high-confidence drug targets, which are druggable, implies that there are still a large number of molecules whose potential as drug targets has not yet been tapped. Structure-based annotation and target prediction provide a wealth of information for TB drug discovery groups to explore.

REFERENCES


Mycobacterium tuberculosis cyclopropane synthetase is required for cording, persistence, and virulence.


ASMscience.org/MicrobiolSpectrum


