Assessing the Progress of *Mycobacterium tuberculosis* H37Rv Structural Genomics

First Author (Corresponding Author):
  Name: Zhuo,
  Surname: Fang\(^1\),
  Email: zf@sun.ac.za
  Tel: +27 21 938 9073
  Fax: +27 21 938 9476

Second Author:
  Name: Ruben,
  Middle Name: Gerhard,
  Surname: van der Merwe\(^1\),
  Email: rvdm@sun.ac.za

Third Author:
  Name: Robin,
  Middle Name: Mark,
  Surname: Warren\(^1\),
  Email: rw1@sun.ac.za

Fourth Author:
  Name: Wolf-Dieter
  Surname: Schubert\(^2\),
  Email: Wolf-Dieter.Schubert@up.ac.za

Last Author:
  Name: Nicolaas
  Middle Name: Claudius,
  Surname: Gey van Pittius\(^1\)
  Email: ngvp@sun.ac.za

\(^1\) DST/NRF Centre of Excellence in Biomedical Tuberculosis Research, US/MRC Centre for Molecular and Cellular Biology, Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, University of Stellenbosch, Francie van Zijl Drive, Tygerberg, 7505, South Africa

\(^2\) Department of Biochemistry, University of Pretoria, Pretoria, South Africa
Abstract

Tuberculosis threatens human health nowhere more than in developing countries with large malnourished and/or immune-compromised (e.g. HIV infected) populations. The etiological agent, *Mycobacterium tuberculosis* (*Mt*b), is highly infectious and current interventions demonstrate limited ability to control the epidemic in particular of drug resistant *Mt*b strains. New drugs and vaccines are thus urgently required. Structural biologists are critical to the TB research community. By identifying potential drug targets and solving their three dimensional structures they open new avenues of identifying potential inhibitors complementing the screening of novel compounds and the investigation of *Mt*b’s molecular physiology by pharmaceutical companies and academic researchers. Much effort has gone into structurally elucidating the *Mt*b proteome though much remains to be done with progress primarily limited by technological constraints. We review the currently available data for *Mt*b H37Rv to extract the lessons they have taught us.

Keywords

*Mycobacterium tuberculosis*, structural genomics, drug target

Abbreviations

All abbreviations in the manuscript are standard in the field.
1. Introduction

With one third of the world population affected, tuberculosis (TB) remains one of the most wide-spread infectious diseases (1). Its etiological agent, *Mycobacterium tuberculosis* (*Mtb*), caused ~1.3 million deaths in 2012 (2). It is partly due to the emergence of multidrug resistant (MDR) and extensive drug resistant (XDR) *Mtb* strains, as well as the complexity of co-infection with human immune-deficient virus (HIV) (2). The importance of identifying new drug targets and developing new drugs cannot be more emphasized. Structural information on *Mtb* proteins can help to characterize known drug targets including *inter alia* arabinosyltransferase and DNA gyrase, respective targets for ethambutol (3) and moxifloxacin (4), and to identify novel molecular intervention strategies (5–8). Understanding the structural effects of resistance mutations can further support the identification of compensating modifications to drugs to stay ahead in the evolutionary race. A comprehensive structural database of *Mtb* proteins is a treasure trove of potential drug targets. This review aims to summarize the current status on structural data for *Mtb* strain H37Rv indicating the progress of *Mtb* structural genomics, providing clues on producing “difficult-to-express” *Mtb* proteins, encouraging international research group collaboration and the exploration on the structures of proteins unique to mycobacteria.

2. Current Status of Structural Data for *Mycobacterium tuberculosis* H37Rv

The first *Mtb* protein structure was deposited with the Protein Data Bank (PDB) in 1994 (9). Since then much effort has been invested in structurally elucidating mycobacterial proteins (10,11). By the end of October of 2014, the PDB contained 1669 mycobacterial protein entries including 1065 without specified strain, 550 for H37Rv, 33 for H37Ra, 10 for CDC1551 and one for F11.

The 550 PDB entries assigned to *Mtb* H37Rv describe 204 distinct proteins and are associated with 251 publications. However, this number dramatically underestimates the true availability of structural data for *Mtb* H37Rv both because many *Mtb* structures in the PDB (potentially including many H37Rv proteins) are not explicitly assigned to any strains and because a close evolutionary relationship of H37Rv to other *Mtb* strains or a more distant kinship to other bacteria means that structural data for H37Rv proteins can be directly inferred. Mapping all 4031 *Mtb* open reading frame (ORF) sequences from Tuberculist (March 2013 update) (12) to all PDB entries using BLAST (13) identifies excellent structural matches for 324 H37Rv proteins (Table S1) using a cut-off of 95 in the “structural match score”, calculated as the product of sequence identity (%) for the best match and the sequence coverage (%) divided by 100. If the cut-off is lowered to 70 (Figure 1), 402 or ~10% of all *Mtb* ORFs may be described as structurally “well defined”. A further 287 ORFs (7.1%) with match values between 30 and 70 provide intermediate to “solid structural data”. For this group, the overall fold or a partial model of good quality can mostly be derived. Overall, good structural data is thus available for ~689 *Mtb* H37Rv ORFs (17.1%). For scores
30, structural data is limited or non-existent implying that structures need to be solved experimentally or that more sophisticated modelling tools may be required.

Figure 1. Distribution of 4031 Mtb H37Rv ORFs by structural match score (sequence identity (%) x coverage (%) / 100) for the best match in the PDB. Most Mtb H37Rv ORFs have no or a low scoring match (≤ 30%)

To avoid ambiguities, the following more detailed analyses will concentrate on those PDB entries explicitly assigned to the Mtb strain H37Rv. This decision is made based on the fact that this laboratory reference strain has been studied and annotated in most detail, that set of proteins is broadly representative of all H37Rv entries, and because the PDB provides a convenient platform for analysis. Although not the focus of this analysis, protein structures from M. smegmatis are often determined in lieu of those from Mtb due to their evolutionary relatedness (14,15).

Structural characterization of Mtb H37Rv proteins mostly involve the elucidation of functional mechanisms by comparing apo-protein structures with complexes involving ions, inhibitors, coenzymes or protein partners bound to the protein of interest. This frequently results in multiple crystal structures for individual proteins and a range of separate publications. While publication output for entries assigned to H37Rv initially increased rapidly from 1998 to 2006, it clearly plateaued thereafter (Figure 2). Apart from characterizing new drug targets, the structural elucidation has also helped to correct mis-annotation of potential protein functions (16–18), to assign functions to previously uncharacterized or “hypothetical” proteins (19–21), and to expand data on Mtb physiology and pathogenicity. Both X-ray crystallography (536 structures, or 97%) and Nuclear Magnetic Resonance Spectroscopy (17 structures, or 3%) were used in solving this set of
protein structures. The quality of the protein structures solved by X-ray crystallography varies (Figure 3)
3. Functional Categorization and Druggability of Mtb H37Rv crystal structures

Ideal anti-TB drug targets either involve enzymes catalysing essential biological reactions or signal transduction components mediating nutrient acquisition or metabolic regulation. Targets should further be unrelated or should share low homology to human orthologs to limit drug toxicity in patients. The 204 structurally characterized proteins from Mtb H37Rv were mostly selected according to these criteria and can be assigned to 5 functional groups: 130 (64%) to “house-keeping metabolism” and “resistance/survival mechanism”, 31 (15%) to “genetic information processing”, 15 (7%) to “cellular information processing and substrate transport”, 11 (6%) “virulence factors”, while 17 (8%) are of unknown function mostly annotated as “hypothetical”. Note that house-keeping metabolism processes dominate amongst studied proteins while only a few virulence associated proteins have been characterized. Structures of previously uncharacterized “hypothetical” proteins were overwhelmingly elucidated by structural genomics consortia. With some exceptions, the structurally solved proteins in H37Rv are mostly similar in function to their recognized orthologs in other organisms and share structural features with other members in the same protein family.

Druggability data of structurally-elucidated proteins is provided by the database TuberQ (22). The druggability scores (DS) fall into four categories including non-druggable (DS ≤ 0.2), poorly druggable (0.2 < DS ≤ 0.5), druggable (0.5 < DS ≤ 0.7), and highly druggable (DS > 0.7) (22). The highest DS for each protein is listed in Table S2 to S14. Of 204 proteins, 18 have not been analysed by TuberQ as yet, one is listed as non-druggable, 16 as poorly druggable, 43 as druggable, and 126 as highly druggable. While this data has been used to develop inhibitors (23,24), no anti-TB drugs currently in clinical trials appears to have been designed from protein structural data (25,26). Instead, drugs were mainly discovered through phenotypic screening of bacteria cells exposed to various compounds (27).

3.1 House-keeping metabolism and resistance/survival mechanisms

Metabolic and environmental stress associated enzymes constitute 64% of available Mtb H37Rv protein structures reflecting the ability of this bacterium to synthesize amino acids, vitamins and cofactors, as well as cell wall components such as peptidoglycan, sugars and mycolic acids. As humans do not code for the latter proteins, these are particularly attractive as drug targets. Enzymes of carbohydrate metabolism, redox homeostasis, nutrient acquisition, oxidative stress relief, antibiotic and harmful chemical elimination share only minimal homology to their human counterparts again making for good potential drug targets. The 126 proteins in this group include nine in carbohydrate metabolism (Table S2), 28 in amino acid metabolism (Table S3), 41 in lipid and polyketide metabolism (Table S4), 15 in nucleotide metabolism (Table S5), 10 in glycan biosynthesis (Table S6), nine in cofactor and vitamin metabolism (Table S7), and 18 in self-defense and detoxification, (Table S8).

3.2 Genetic information processing
Genetic integrity is vital for survival. Gene transcription and protein production are hence critical metabolic processes. Targeting components in these pathways will kill or limit the growth of \textit{Mtb}. Protein structures assigned to this category include 17 proteins in transcriptional regulation (Table S9), five in translation/protein folding (Table S10) and nine in DNA replication/repair (Table S11).

### 3.3 Cellular information processing and substrate transport

Cellular signalling pathways allow bacteria to sense and respond to environmental stimuli such as chemical gradients including \(\text{PO}_4^{3-}\), \(\text{O}_2\), \(\text{NO}\), \(\text{Fe}^{2+}\), \(\text{N}_2\), cAMP. Proteins in this group mostly belong to two-component systems and serine/threonine kinases (PknB) though their exact function often remains unclear. Fifteen protein structures fall into this category (Table S12).

### 3.4 Virulence factors and hypothetical proteins

Despite the molecular basis for \textit{Mtb} pathogenicity being of critical importance, knowledge in this area remains quite limited. \textit{Mtb} virulence factors appear to mostly be directed towards ensuring mycobacterial survival in human macrophages. Only 11 of 203 or 6% of \textit{Mtb} H37Rv structurally elucidated proteins belong to this group (Table S13). Identifying human targets of virulence factors as well as crystallizing the associated molecular complexes remains challenging with the result that data in this regard is currently not available.

Structural analysis of 17 proteins of unknown function (Table S14) has led to novel fold discovery and the assignment of possible functions to some proteins. Thus, Rv2714 was found to be a representative of a new nucleoside phosphorylase-like family of actinobacterial proteins (28); Rv1980c represents a new family of secreted \(\beta\)-grasp proteins (29); Rv2827c has a unique DNA-binding fold although the exact function remains unknown (30); Rv1155 resembles flavin mononucleotide (FMN)-binding proteins but has no affinity towards FMN (31). Rv0802, resembling members of the GCN5-related N-acetyltransferase family, utilizes succinyl-CoA as substrate rather than acetyl-CoA, the first in this family (32). In contrast to the above, no function could be identified for many structurally-elucidated hypothetical proteins. Examples include Rv0020c, a substrate of Ser/Thr protein kinase PknB (33), and Rv2140c with marked homology to phosphatidylethanolamine-binding proteins indicating a potential role in lipid metabolism (34) (Table S13). These highlight the important contribution structural biology can make to our understanding of mycobacterial biology.

### 4. Technical aspects of \textit{Mtb} structural analysis

Structural elucidation of \textit{Mtb} proteins remains challenging mainly due to difficulties with gene expression, protein solubility, protein purification and crystallization. Unexpectedly, codon-optimization of \textit{Mtb} genes has not been frequently applied. With costs for DNA
synthesis steadily decreasing this modification may become more accessible in future. Overall, affinity tag fusion proteins remain popular: 172 of 203 proteins or 85% were produced with a tag, though tagless proteins also feature (31 or 15%). N-terminal His$_6$-tags dominate (56%), followed by C-terminal His$_6$ (23%). Larger tags such as glutathione S-transferase (GST, 26 kDa), maltose-binding protein (MBP, 42.5 kDa) and intein/chitin-binding protein (CBP, 55 kDa) which may help to increase solubility were used less frequently. The GST-tag was correspondingly used for six proteins (3%), MBP for four (1%), while CBP and streptavidin tag were used once each (0.5%).

Proteins used to derive the deposited *Mtb* H37Rv structures were with very few exceptions all heterologously produced in *Escherichia coli* (536 of 553 PDB entries or 200 of 203 unique proteins, See Table 1). Established protocols and the superior efficiency of this host compared to *Mtb* account for this landslide. The *Mtb* genome is more G/C rich (>65%) and its codon usage differs from that of *E. coli*. It incorporates a higher proportion of glycine, alanine, proline and arginine and *Mtb* post-translational modification machinery is lacking in *E. coli* (35,36). *M. smegmatis*, a fast-growing saprophytic relative of *Mtb*, is a potential expression host for *Mtb* genes overcoming many of the restrictions of *E. coli* (36–39). Unexpectedly, however, only nine proteins of H37Rv represented by 27 structures were produced in *M. smegmatis* (37,40–46). Clearly the faster growth and production rates of *E. coli* ensure that it remains the most popular expression host. Also strikingly, BL21 (DE3) was used most frequently to express *Mtb* genes despite the presence of rare codons. Only a small proportion of *Mtb* proteins were produced in Rosetta or CodonPlus strains that code for tRNAs of these rare codons. Again the observed bias towards BL21 (DE3) may largely be practical with researchers simply using the most widely available strain. In future, other expression strains may come to the help for non-standard proteins such as membrane proteins.

Table 1. Expression host for H37Rv protein production

<table>
<thead>
<tr>
<th>Expression Host</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> BL21(DE3) and its derivative strains</td>
<td>159</td>
</tr>
<tr>
<td><em>E. coli</em> Rosetta or BL21 CodonPlus strain</td>
<td>23</td>
</tr>
<tr>
<td><em>E. coli</em> B834</td>
<td>7</td>
</tr>
<tr>
<td><em>M. smegmatis</em> mc$^2$ 155</td>
<td>7</td>
</tr>
<tr>
<td><em>E. coli</em> C41(DE3)</td>
<td>4</td>
</tr>
<tr>
<td><em>E. coli</em> DH5α</td>
<td>3</td>
</tr>
<tr>
<td><em>E. coli</em> Origami</td>
<td>2</td>
</tr>
<tr>
<td><em>Pseudomonas putida</em> KT4224</td>
<td>2</td>
</tr>
<tr>
<td><em>M. smegmatis</em> mc$^2$ 4517</td>
<td>2</td>
</tr>
</tbody>
</table>
Production conditions were optimized for most proteins. Overwhelmingly lac promoter constructs (T7 and tac) were used for isopropyl β-D-1-thiogalactopyranoside (IPTG) induction in Lysogeny Broth (183 of 203 proteins). Only for three proteins was Terrific Broth (TB) used, auto-inducing media for 11, and arabinose induction for six proteins. Plotting production temperature against frequency (Figure 4) indicates that production temperatures between 16 and 25°C were most common, with additional spikes at 37°C and 30°C. Clearly, lower temperatures limit heat shock protein production, proteolytic degradation, protein aggregation (inclusion bodies) and improve protein stability (47). Though only rarely reported, E. coli strains such as Arctic Express and expression vectors such as pCOLD, co-produce cold shock chaperone proteins which help to overcome production hurdles for Mtb proteins. Arylamine N-acetyltransferase (TBNAT), for example, was produced insolubly in E. coli BL21 (DE3) pLysS but solubly in ArcticExpress (35).

![Figure 4](image.png)

Figure 4. Protein production temperatures used for Mtb H37Rv proteins in E. coli. The number of proteins is plotted against production temperature. Most proteins were produced at moderately low temperatures (16-25°C).

Protein constructs resulting in successful structures are mostly small to medium in size (10-60 kDa) with relatively neutral pI (4-8) (Figure 5). The distribution pattern of those structures on MW vs pI plot does not deviate dramatically from that of all H37Rv ORFs (Figure 5). A majority of proteins are further inherently soluble and stable enzymes or signal transduction
components. These features clearly enhance successful production in *E. coli* but indicate that future progress will presumably be much slower. Membrane protein structures are correspondingly rare with structural studies mostly being limited to soluble domain such as in FtsX (48), MmpS4 (49), MmpL3 and MmpL11 (50).

![Figure 5](image.png)

Figure 5. A plot of pI versus molecular weight of all *Mtb* H37Rv protein structures in the PDB and all *Mtb* H37Rv ORFs.

Insolubility of target *Mtb* proteins remains a major hurdle for structural analysis. The solubility of secreted target proteins may be improved by co-expressing the soluble domain of the membrane-anchored disulphide bond-forming (Dsb) protein (Rv2969c) (51,52). Similarly, GlgB (Rv1326c), Cyp125A1 (Rv3545c), and tuberculosinyl transferase (Rv3378c) were converted to soluble proteins by co-producing *E. coli* chaperone proteins (17,53,54). Corresponding *Mtb* chaperones such as Cpn60.1, an ortholog of *E. coli* GroEL, significantly improved the solubility of DprE1, the target for the drug benzothiazinone (55). Protein solubility may also be optimized by co-expressing other proteins of the same metabolic pathway. Siroheme-dependent sulfite reductase (Rv2391) was hence solubly produced by co-producing uroporphyrinogen III C-methyl-transferase (Rv2847c) (56). Similarly, non-ribosomal peptide synthetases, MbtE and MbtF of mycobactin biosynthesis, were solubly produced only by co-producing 8 kDa MbtH with a previously unknown role in the pathway (57,58). Yet another strategy of obtaining structural data for difficult or insoluble *Mtb* proteins is to replace them with mycobacterial orthologues. Thus, Mycosin-1 and Mycosin-3
proteases from ESX-1 and ESX-3 Type VII secretion systems were structurally characterized from *M. smegmatis* in place of their *Mtb* counterparts. Amino acid sequence identities of 72% and 60%, respectively ensure that structural features may directly be inferred for the latter (14,15,59). Underlying causes for difficulties in producing *Mtb* proteins mostly remain unresolved as projects are generally discontinued if problems prove intractable.

5. **Scope of future structural genomics studies**

In choosing targets for structural analysis, proteins unrelated to existing PDB entries may be of most interest as they may represent new protein families or new functions. Rv2179c, for example, discovered to have RNase activity but without structural similarity to other RNases, was identified as the founding member of a large family of bacterial RNases (19). Many proteins without structural counterparts in the PDB, however, are membrane proteins limiting rapid progress in structural analysis. Furthermore some *Mtb* H37Rv ORFs code for highly unusual proteins. Rv3512 for example contains 598 glycines (55%) and 140 alanines (13%) within 1079 amino acids, identifying this protein as potentially inherently unfolded.

6. **Structural Biology Research Group World Wide**

Table S15 lists the principle investigators, the institutions, the countries and the category of proteins analysed. Overall, 118 principle investigators contributed to the 252 publications for *Mtb* H37Rv protein structures in the PDB. Of these, 59 involved with more than one publication are listed. Groups are mostly based in Europe and Northern America but complemented by groups from India, New Zealand, China and South Korea.

7. **Acknowledgement**

We would like to thank Samantha Sampson for invaluable discussion and comments.

8. **Author Contribution**

ZF, WDS and NCGP conceptualized the project; ZF, RGM and WDS did the bioinformatical data analysis; ZF, RW, WDS and NCGP wrote the review.

9. **Conflict of Interest**

The authors declare no conflict of interest.
Reference


2. WHO, Global tuberculosis report 2013. WHO.


