

Using Structural Genomics to Understand *Mycobacterium tuberculosis*

J. Shaun Lott

Laboratory of Structural Biology and Centre for Molecular Biodiscovery,
School of Biological Sciences, University of Auckland, New Zealand

Mycobacterium tuberculosis, the causative agent of TB, is the most devastating of all human pathogens. For roughly 10,000 years of human history, tuberculosis (also known as consumption or the white plague) has been a constant cause of mortality. It was the major cause of death in Europe in the 17th to 19th centuries, and even today more people are killed worldwide by *M. tuberculosis* infection than by infection with any other bacterium. Recent estimates of the global problem indicate that more than 1,000 new cases of TB occur every hour, resulting in the death of more than 7,000 people per day (1,2). The global TB problem was recognised as sufficiently serious to be declared the first global health emergency by the World Health Organisation in 1993. The incidence of TB continues to rise in the developing world as a result of a number of factors, such as the HIV pandemic and declining social standards caused by rapid population growth and urbanisation. Compounding this problem is the emergence of TB strains that are resistant to one or all of the preferred antibiotics that would otherwise allow successful treatment of the disease. Significantly, there is no room for complacency in the developed world, as the incidence of TB in the major industrialised nations has increased throughout the 1990s, reversing the steady decline of previous decades (3).

Robert Koch first isolated the tubercule bacillus in 1882, and showed that it was the agent responsible for human disease (4). Despite being studied since the infancy of microbiology, the pathogenic mycobacteria have remained relatively poorly understood at the molecular level, largely as a result of their intractability. *M. tuberculosis* grows so slowly (doubling once every 24 hours under optimal growth conditions) that it takes 3-4 weeks to form a colony on a plate, and *M. leprae* (the causative agent of leprosy) remains unculturable *in vitro*.

Despite a lack of knowledge about many fundamental aspects of mycobacterial physiology, the 20th century saw large advances in the treatment of mycobacterial disease. The pioneering work of Albert Calmette and Camille Guérin (5) led to the development of the Bacille Calmette-Guérin (BCG) vaccine (an avirulent strain of *M. bovis*, the bacterium that causes bovine TB). This has been used to immunise many millions of people, with efficacy varying from 10% to 80%, depending on the population in question. The development of antimicrobial chemotherapy in the 1950s led to spectacular improvement in the treatment of infected individuals, but was quickly followed by the realisation that single-drug therapy merely encouraged the outgrowth of drug-resistant strains of the bacillus. Current antibiotic treatment therefore consists of a combination of up to four antibiotics for a period of 6-9 months – this is ‘short course’ therapy! Multi drug resistant (MDR) strains of TB are defined as those that are resistant to at least two of the primary drugs used in this therapy. Treatment of MDR-TB is extremely costly and limited in its success: in many parts of the world, and for immune compromised individuals, this makes it essentially untreatable. Thus,

despite the advances in treating TB that the 20th century has brought, there is a strong need for new therapeutic avenues at the start of the 21st century.

Mycobacterial genomics

The complete sequence of the genome of the virulent H37Rv strain of *M. tuberculosis* was completed in 1998 (6). The GC-rich genome is approximately 4.4 megabases in size, and contains roughly 4,000 open reading frames (ORFs). Despite an ongoing program of re-annotation of the genome sequence, 4 years after its completion still only half of the ORFs could be assigned even a putative function (7). Thus, it appears that roughly half of mycobacterial biology is hidden from our current knowledge (8). That said, identifiable ORFs in the genome sequence give many insights into mycobacterial physiology.

One striking feature is the strong propensity for lipid metabolism: more than 8% of the genome is dedicated to this activity. Moreover, at least 100 lipolytic enzymes have been identified, something not seen in any other bacterium to date. Another unusual element of the genome is the presence of multigene families of PE and PPE proteins, named after characteristic proline-glutamate motifs, which again make up about 8% of the coding potential of the genome. Their precise functions are still unclear, but evidence is growing that they are variable cell wall proteins with structural roles. Indeed, some of them may be determinants of pathogenic virulence. Identifiable virulence factors have classically been lacking in TB pathology (9), but the genome sequence contains many candidate genes. These include homologues of virulence factors secreted from *Listeria*, secreted phospholipases (10) and putative cytolysins (11), all of which may prove to have a role in the ability of the bacterium to survive engulfment by macrophages.

Post-genomic analysis

The availability of the *M. tuberculosis* genome sequence has enabled a variety of post-genomic techniques to be employed to understand the bacterium’s hidden half. For instance, DNA microarray experiments have been used to discover which genes are upregulated by cell wall-disrupting antibiotics (12) and by hypoxia (13). The adaptation to this is thought to be central to the bacterium’s characteristic ability to remain latent in asymptomatic patients for long periods of time – in some cases many years.

Comparative genomic studies between virulent and avirulent strains of mycobacteria have uncovered several regions of difference (RD) between otherwise genetically very similar organisms (14,15). These regions are useful both as diagnostic markers to rapidly distinguish between strains, and as clues to the origins of virulence. The presence or absence of RD1, for example, appears to correlate with the ability of the bacterium to cause disease (16). However, as the proteins encoded by this region are functionally uncharacterised, more work is required to understand this observation in molecular terms.

More subtle comparative genomics experiments are also becoming possible, such as 2-dimensional bacterial genome display. This method uses high-resolution electrophoresis to separate digested genomic DNA on the basis of size and G+C content; it is able to differentiate between isogenic strains that differ only by a single frame-shift mutation (17).

M. tuberculosis has classically been difficult to work with genetically, and the generation of knockouts and defined mutants has been difficult due to the bacterium's propensity for illegitimate recombination over homologous recombination. However, recent technical advances have provided the ability to produce specific knockout mutants (18). Guided by the roadmap that the genome sequence provides, these technical advances are now coming into their own in identifying the role of many genes in bacterial pathogenesis and intracellular survival.

Structural genomics

The term 'structural genomics' encompasses a variety of approaches whose ultimate goal is to provide 3-dimensional structures for all of the proteins encoded by genomes. This ambitious goal includes the use of genomic information to guide protein structure discovery that may be applied, for example, to facilitate drug development. It also includes the converse, namely, the use of protein structure analysis to add value to genomic sequence data by deducing function from structure. For either approach to be effective, large-scale (high throughput) methods for rapid structure determination must be developed.

The TB Structural Genomics Consortium (<http://www.doe-mbi.ucla.edu/TB>) is an international collaboration of research groups with the common aim of using protein structure both to understand TB biology and to facilitate the development of new anti-TB drugs. Funded by the US National Institutes of Health, these groups have established central facilities for the following tasks (US locations in brackets):

- high throughput gene cloning, expression and purification of gene products (Los Alamos National Laboratory)
- high throughput protein crystallisation (Lawrence Livermore National Laboratory)
- advanced structure determination methods (Los Alamos National Laboratory, Texas A&M, University of California, Berkeley)
- synchrotron data collection (Brookhaven National Laboratory)
- bioinformatics and database development (University of California)
- gene knockouts (Albert Einstein College of Medicine)

These facilities, and the technology and materials they develop, are available to all participants within the Consortium. Operationally, members of the Consortium choose their own targets for structural analysis and communicate progress through a shared web-accessible central database. Where two groups are interested in the same protein they are encouraged to communicate and collaborate. This open and coordinated approach is intended to avoid duplication

and maximise the coverage of TB biology. A second advantage of this approach is that members share the tools, sometimes specific to TB proteins, that they develop, and can share information on current methodology.

The Laboratory for Structural Biology at the University of Auckland is one of the founding members of the Consortium, and has focused on solving the structures of proteins in the following thematic areas:

Biosynthesis of amino acids shown to be essential for mycobacterial growth *in vivo*, e.g. leucine (19), tryptophan and proline (20). The structure of LeuA (2-isopropylmalate synthase) has recently been solved in this laboratory by Nayden Koon and Chris Squire. This enzyme catalyses the first step in leucine biosynthesis, and is a dimeric protein composed of two domains. The catalytic domain is a classic $(\alpha/\beta)_8$ barrel, whereas the regulatory domain (Fig. 1) shows structural similarity to double-stranded RNA binding proteins. Analysis of the protein structure has already revealed that the enzyme is zinc-dependent and has identified a likely active site.

Biosynthesis of cofactors that may be important under hypoxic conditions, e.g. menaquinone. The structures of the menaquinone biosynthetic enzymes MenG and MenB have been solved by Jodie Johnston and Vic Arcus. The structure of the *menG* gene product (Fig. 1) has revealed a misannotation in the genome sequence: *menG* is annotated as SAM-dependent 2-



Fig. 1. Examples of protein structures from *M. tuberculosis* solved in Auckland.

Left: The regulatory domain of LeuA, a key enzyme in branched-chain amino acid biosynthesis. (Structure solved by Nayden Koon and Chris Squire).

Top right: The structure of MenG, a protein incorrectly annotated as a methyltransferase. (Structure solved by Jodie Johnston and Vic Arcus).

Bottom right: The structure of Rv1347c, a putative aminoglycoside acetyltransferase. (Structure solved by Graeme Card).

Cartoons produced using PyMol (www.pymol.org).

demethylmenaquinonemethyltransferase based on sequence similarity to the *Escherichia coli* enzyme. This is the final step in menaquinone biosynthesis. The fold of MenG is quite different from that of any known methyltransferase, and instead resembles that of known phosphotransferases. From the structure, it is clear that this enzyme has probably been misannotated (21), with this reaction instead being performed by the UbiE protein. The actual function of MenG remains to be determined.

Potential antibiotic resistance factors, and proteins upregulated by antibiotic treatment (12). The gene Rv1347c is annotated as a possible aminoglycoside 6'-N-acetyltransferase, a potentially aminoglycoside antibiotic inactivator. The structure of Rv1347c confirms that it is indeed a member of the GCN5-related N-acetyltransferase (GNAT) family of aminoglycoside N-acetyltransferases (AAT), although its exact substrate remains to be determined. The *ini* operon (Fig. 2) has been shown to be upregulated in response to the antibiotic isoniazid (12) and also in response to other cell wall-disrupting antibiotics, suggesting that these genes may be involved in maintaining cell wall integrity under conditions of cellular stress. Sequence conservation between IniA and IniC and also between IniB and IniD indicate the presence of shared domains, which we have confirmed by limited tryptic digestion (Fig. 2). In our lab, Moyra Komen is currently close to solving the structure of the conserved domain from IniD using NMR; she has grown diffracting crystals of the conserved domain

from IniC. Despite having a putative physiological role within the bacterium these proteins have not yet been assigned a biochemical function, as they have no homologues identifiable by sequence similarity. Hence, structural characterisation may be the only way to gain an insight into their function.

Concluding remarks

The completion of the sequence of the *M. tuberculosis* genome was the starting point for a new genomic phase of investigation into mankind's oldest known pathogen. The TB Structural Genomics Consortium has set out to solve 400 protein structures in 5 years, focussing on proteins with potential for use in the development of new drugs and vaccines. Technological developments are transforming the speed with which new protein structures can be determined. Thus far, the Consortium as a whole has solved the structures of almost 30 proteins from *M. tuberculosis*. Many of the protein structures, chosen for their roles in biosynthetic pathways, growth or virulence, will provide the templates for new drug discovery. Others will illuminate TB biology by the discovery of new structural, evolutionary and metabolic relationships, the clarification of putative functions and the suggestion of possible functions for some of the many unknown entities that have been identified in the complete genome sequence.

References

- Rattan, A., Kalia, A., and Ahmad, N. (1998) *Emerging Infectious Diseases* **4**, 195-209
- World Health Organisation. (2000)
- Stokstad, E. (2000) *Science* **287**, 2391
- Koch, R. (1882) *Berl. Klin. Wochenschr.* **19**, 221-230
- Calmette, A., Guérin, C.L.N., and Boquet, A. (1927) *Ann. Inst. Pasteur* **XLI**, 201-232
- Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., et al. (1998) *Nature* **393**, 537-544
- Camus, J.C., Pryor, M.J., Medigue, C., and Cole, S.T. (2002) *Microbiology* **148**, 2967-2973
- Lott, J.S., and Baker, E.N. (2003) *New Zealand Science Rev.* **60**, 43-47
- Young, D.B. (1998) *Nature* **393**, 515-516
- Songer, J.G. (1997) *Trends Microbiol.* **5**, 156-161
- Wren, B.W., Stabler, R.A., Das, S.S., Butcher, P.D., Mangan, J.A., Clarke, J.D., Casali, N., Parish, T., and Stoker, N.G. (1998) *Microbiology* **144**, 1205-1211
- Wilson, M., DeRisi, J., Kristensen, H.H., Imboden, P., Rane, S., Brown, P.O., and Schoolnik, G.K. (1999) *Proc. Nat. Acad. Sci. USA* **96**, 12833-12838
- Sherman, D.R., Voskuil, M., Schnappinger, D., Liao, R., Harrell, M.I., and Schoolnik, G.K. (2001) *Proc. Nat. Acad. Sci. USA* **98**, 7534-7539
- Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., and Small, P.M. (1999) *Science* **284**, 1520-1523
- Gordon, S.V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K., and Cole, S.T. (1999) *Mol. Microbiol.* **32**, 643-655
- Cole, S.T. (2002) *Microbiol.* **148**, 2919-2928
- Dullaghan, E.M., Malloff, C.A., Li, A.H., Lam, W.L., and Stokes, R.W. (2002) *Microbiol.* **148**, 3111-3117
- Jacobs, W.R., Barletta, R.G., Udani, R., Chan, J., Kalkut, G., Sosne, G., Kieser, T., Sarkis, G., Hatfull, G., and Bloom, B.R. (1991) *Methods Enzymol.* **204**, 537-555
- Hondalus, M.K., Bardarov, S., Russell, R., Chan, J., Jacobs, W.R., and Bloom, B.R. (2000) *Infect. Immun.* **68**, 2888-2898
- Smith, D.A., Parish, T., Stoker, N.G., and Bancroft, G.J. (2001) *Infect. Immun.* **69**, 1142-1150
- Johnston, J.M., Arcus, V.L., Morton, C.J., Parker, M.J., and Baker, E.N. (2003) *J. Bacteriol.*, in press

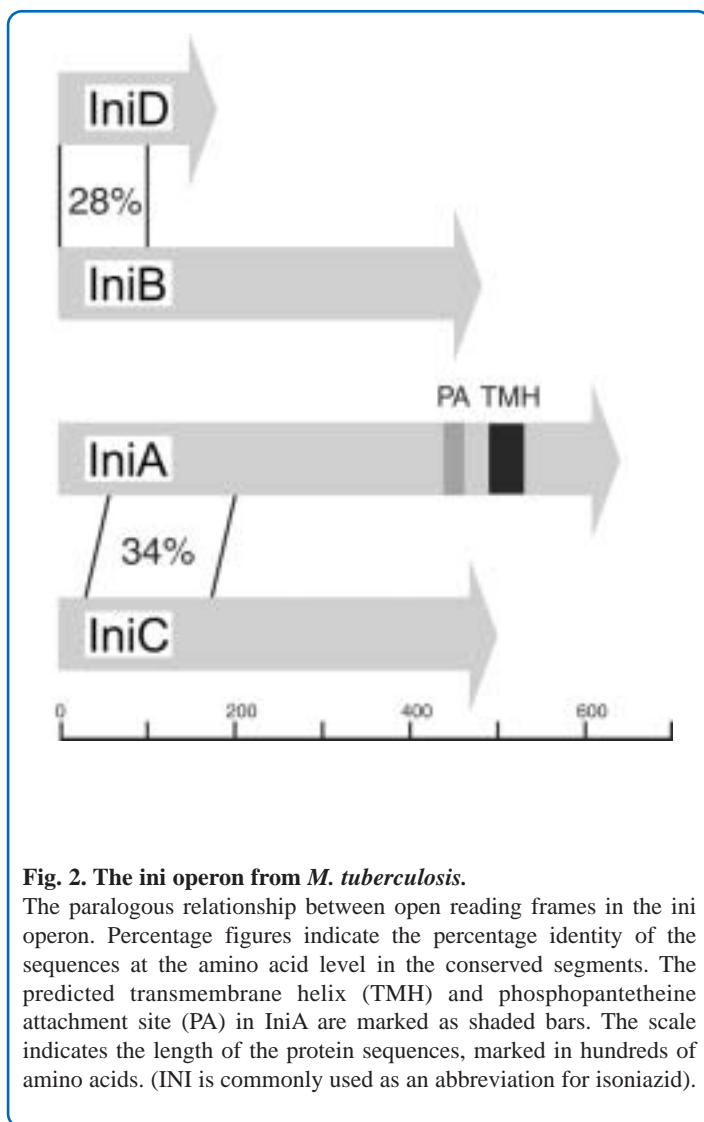


Fig. 2. The *ini* operon from *M. tuberculosis*.

The paralogous relationship between open reading frames in the *ini* operon. Percentage figures indicate the percentage identity of the sequences at the amino acid level in the conserved segments. The predicted transmembrane helix (TMH) and phosphopantetheine attachment site (PA) in IniA are marked as shaded bars. The scale indicates the length of the protein sequences, marked in hundreds of amino acids. (INI is commonly used as an abbreviation for isoniazid).

Fig. 1



Fig. 2

