Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Structural genomics approach to drug discovery for *Mycobacterium tuberculosis*

Thomas R Ioerger[1] and James C Sacchettini[2]

Structural genomics has become a powerful tool for studying microorganisms at the molecular level. Advances in technology have enabled the assembly of high-throughput pipelines that can be used to automate X-ray crystal structure determination for many proteins in the genome of a target organism. In this paper, we describe the methods used in the Tuberculosis Structural Genomics Consortium (TBSGC), ranging from protein production and crystallization to diffraction data collection and processing. The TBSGC is unique in that it uses biological importance as a primary criterion for target selection. The over-riding goal is to solve structures of proteins that may be potential drug targets, in order to support drug discovery efforts. We describe the crystal structures of several significant proteins in the *M. tuberculosis* genome that have been solved by the TBSGC over the past few years. We conclude by describing the high-throughput screening facilities and virtual screening facilities we have implemented for identifying small-molecule inhibitors of proteins whose structures have been solved.

Addresses
[1] Department of Computer Science and Engineering, Texas A&M University, United States
[2] Department of Biochemistry and Biophysics, Texas A&M University, United States

Corresponding author: Sacchettini, James C (jim.sacchettini@gmail.com)

Current Opinion in Microbiology 2009, **12**:318–325

This review comes from a themed issue on
Techniques
Edited by Jerome Garin and Grant Jensen

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2009.04.006

Introduction

Structural genomics is a route to understanding microbial organisms at a molecular level. Structural genomics refers to large-scale efforts to determine as many of the unique structures of proteins in an organism as possible, primarily through X-ray crystallography. It complements genome sequencing and other technologies like DNA microarrays for assessing gene expression. Whole-genome sequence comparisons can provide information about gene conservation, operon structure, duplication/loss, synteny, and evolutionary relationships. However, many genes in bacterial genomes are unannotated (lacking sufficient homology to known families), half of all protein families lack a structural representative, and many metabolic pathways remain incomplete, hampering our efforts at understanding phenomena from basic metabolism, to environmental sensing and response, to host–pathogen interactions. Thus, by determining unannotated protein structures, unexpected family relationships can often be discovered, leading to hypotheses about their potential functions that can be tested.

Structural genomics was born from advances in technology that made high-throughput structure determination possible. With the advent of high-intensity beam-lines at synchrotrons and new phasing techniques such as multi-wavelength anomalous dispersion (MAD) [1], (and more recently SAD [2] and SIR), along with more powerful computational algorithms for data processing, refinement and model building [3–6], high-resolution structures can be solved more systematically, often within days of obtaining crystals. When this is coupled with new methods for expression, purification, and crystallization [7•], a high-throughput pipeline can be assembled for solving the structures of many targets in a rapid and automated way.

In the late 1990s, several approaches were proposed for applying these new high-throughput structure-determination pipelines, collectively called Structural Genomics. One approach was to use it to fill out 'fold-space' [8]. It has been recognized early on that, as more structures were being solved, many tended to fall in existing fold classes. This was true even for unexpected cases where the sequence homology to other members in the fold family was low (in 'twilight zone', below threshold of statistical significance for detecting relationship). Early estimates were of about 1000 total folds [9], though this has crept up slightly to ~10 000 folds [10]. Since significant homology to a known fold family is often sufficient to broadly characterize its function, one way to use structural genomics is to systematically solve the structures of only those proteins without any detectable homology to any currently known fold family [11], with the hope that all remaining folds will eventually be sampled. This can then be used to establish a more-or-less complete library of folds, against which genes sequences can be analyzed by new, more sensitive fold detection methods, such as Hidden Markov Models [12] or sequence-structure threading [13].

An alternative approach to using high-throughput SG pipelines is to focus on solving structures of functionally important, biologically interesting, or medically relevant targets. Within a given genome, ORFs can be prioritized by many different criteria. For the study of infectious organisms, a particular driving force is drug discovery [14–15]. Here the pathways of interest are those that are essential to survival (or virulence) of the organism, such that inhibiting members on the pathway will lead to cidality. One of the most effective approaches to inhibitor discovery is structure-based drug design [16], where crystal structures of proteins, especially in complex with substrates, substrate analogs, or transition-state mimics, which can be exploited to design compounds that achieve with higher affinity through specific hydrophobic, polar, and electrostatic interactions in the active site. This approach has been used to rationalize and improve a number of antibacterial agents [17].

Around 1999–2001, a number of Structural Genomics programs were initiated by various funding agencies around the world. This included the NIH Structural Genomics Centers (SGCs) funded through the Protein Structure Initiative (PSI; http://www.nigms.nih.gov/Initiatives/PSI), as well as several other SGCs in other countries around the world (e.g. Northwest SGC in the UK, Protein Structure Factory in Germany, RIKEN in Japan). Each Center had a different focus or emphasis, and each developed their own high-throughput pipelines with different strengths and technologies. Most SGCs selected model organisms defining genomes from which to choose structures to solve, including several microbial representatives. The Joint Center for Structural Genomics aimed to solve structures of widely conserved proteins across all domains of life, including *Thermotoga maritima*, a hyperthermophile, as a bacterial representative. The Southeast Center for Structural Genomics chose to select targets from the archaebacterium *Pyrococcus furiosus* for comparative purposes. The focus of the Midwest Structural Genomics Consortium was on solving structures of genes in organisms causing infectious diseases, including *B. anthracis*, *Y. pestis*, and *Salmonella*. Finally, the Center for Structural Genomics of Pathogenic Protozoa, focused on target selection from *Plasmodium*, *Leishmania*, and *Trypanosoma* species, each also associated with important human diseases.

Over the past decade, great progress has been made. Collectively, over 3300 protein structures have been contributed to the RCSB through PSI-sponsored structural genomics consortia since 2000, and 20% of all new folds had been discovered in the process, with nearly equal contributions by other non-US SGCs around the world [18••]. Bottlenecks still remain. Some proteins remain extremely difficult to express in soluble form or crystallize, and seem to defy even the newest techniques available. There is a continual push to solve the structures of larger/multi-domain proteins, and recently focus has shifted to solving the structures of protein complexes [19••] and membrane proteins [20], where many important interactions happen. Nonetheless, the gradual accumulation of 3D protein structures from the genomes of various microbial organisms will have a significant impact on our understanding of their biology, as well as paving the way for the discovery of new drugs.

In the remainder of this review, we focus on the contribution of structural genomics to our understanding of tuberculosis, the structure determination pipeline implemented by the Tuberculosis Structural Genomics Consortium, and target selection based on potential relevance to drug discovery as a motivating goal.

The Tuberculosis Structural Genomics Consortium

The Tuberculosis Structural Genomics Consortium (TBSGC) was established in 2000, with centralized core facilities serving over 100 collaborating research labs around the world. Rather than targeting low-homology (unannotated) proteins to fill-out ‘fold-space’, the TBSGC has instead applied the concept of a high-throughput pipeline to determining the structures of functionally important proteins [21,22•], to improve our understanding of metabolic pathways, and ultimately to facilitate the drug discovery process.

Tuberculosis, an infection caused by the obligate human pathogen *Mycobacterium tuberculosis*, typically at pulmonary sites, remains a world-wide health crisis, causing nearly two million deaths per year, with an estimated one-third of the human population carrying a latent infection [23]. Drug treatment is a laborious process, requiring daily dosage of two to four drugs over six months, and compliance is poor. Most recently, there has been an alarming rise of multi-drug resistant and extensively drug resistant TB world-wide [24••], making development of new drugs crucial. Progress has been made in understanding the basis of infection, including interaction with macrophages, host immune response, and transition to a persistent state [25,26]. This has been facilitated by new tools, ranging from laboratory models of persistence (hypoxia, nutrient starvation, reactive nitrogen species), animal models, and DNA microarrays for studying changes in gene expression patterns. Nonetheless, a number of fundamental questions about the biology of TB remain unanswered, and it has been over 30 years since the approval of any new drugs for chemotherapy.

To date, crystal structures for 257 out of the ~4000 genes in the *M. tuberculosis* genome have been determined and deposited in the Protein DataBank (PDB) (118 were contributed by TBSGC members). Of the remaining ones, 40% are unannotated (labeled as ‘hypothetical proteins’). In some cases, structure determination has also revealed mistakes in annotation, as with *menG* (Rv3853), originally expected to be a menaquinone methyltransferase, but shown later to be a member of the *RraA* family, a regulator of RNase E [27,28].

The TBSGC has implemented a high-throughput structure determination process by X-ray crystallography that exploits a number of different technologies [22•]. Most of the relevant 3989 ORFs annotated in the TB genome have been cloned into the Gateway system (Invitrogen Inc.), which affords a rapid and convenient method for recombination into expression systems [29•]. For expression and purification, the Consortium has a protein-production core facility at Los Alamos National Lab utilizing robotics, cell-free expression systems and high-throughput solubility assays [30]. Some labs use traditional techniques for trying to increase the solubility of recalcitrant proteins, such as making fusions with MBP (maltose-binding protein) or truncations. Some labs have experimented with directed evolution [31], as well as replacing surface residues using predictions from the surface-entropy server [32•] to increase solubility. Purification is most routinely achieved by attaching an N-terminal His6 tag with a TEV cleavage site, for separation on a Ni-column and then removal of the fragment via proteolysis. While some labs use commercially available screens or custom-developed sets of solvents/conditions for crystal screening, a randomized method for screening has been developed for efficiently surveying a broader set of conditions [33]. Also, recent advances in micro-fluidic technology have enabled crystallization of some proteins on 'chips' using free interface diffusion in the facility at UC, Berkeley [34]. The TBSGC uses several high-intensity synchrotron beamlines for collecting X-ray diffraction data, including ALS (Stanford) and APS (Chicago). Most structures of late have been solved via multi-wavelength anomalous dispersion (MAD; [1]), or by molecular replacement [6], though some by SAD or SIR. Crystallographic software systems such as CCP4 and Phenix [35] are used for data manipulation, phasing, automated model building, and refinement.

Given that the mission of the TBSGC is based on defining the structures of new drug targets, targeting is not as straightforward a process as with other structural genomics programs. In the TBSGC, the method we use is based on a bioinformatics approach, by combining as much relevant data as possible to prioritize targets in terms of the likelihood that they will be good drug targets. This approach (following [36]) takes into account all available data on drugability, enzyme pathway analysis, essentiality (e.g. via transposon knockouts [37]), and gene expression under different models of persistence to identify genes whose inhibition might lead to bacterial cell death (or attenuation of virulence). An interactive web service called Target Explorer has been implemented to allow investigators to experiment with dynamically adapting the weighting of different criteria, including multiple DNA micro-array datasets, for selecting preferred targets for structure determination. This is one of several informatics tools provided on the Consortium web site for information sharing, http://webtb.org.

Recently solved structures

The structures of a number of interesting genes in the TB genome have been solved over the past few years that contribute to our understanding of the biology of this organism, and open new avenues for drug discovery. Here we describe several for their potential as drug targets.

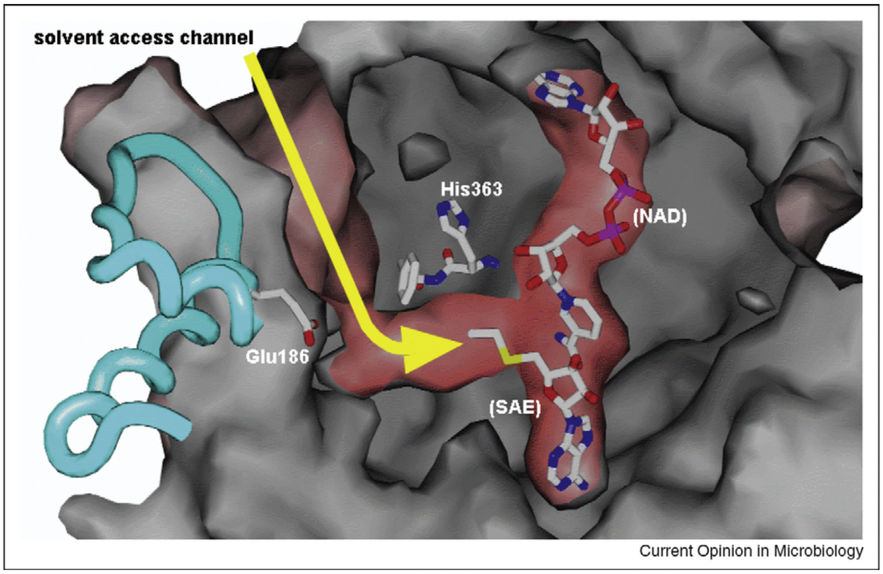
Anthranilate phosphoribosyltansferase (*trpD*, Rv2192c) is the first step in the tryptophan biosynthesis pathway, and appears to be essential for virulence in mice [38]. Although a complex with anthranilate has not yet been solved, comparison between the apo structure and the complex with PRPP (PDB: 2bpq, 1zvw; [39]) show significant conformational changes in the form of hinge-bending that brings domains together.

S-adenosyl homocysteine hydrolase (*sahH*, Rv3248c) plays an important role in maintaining the intracellular balance between co-factors S-adenosyl methionine (SAM), used for a wide variety of methylation reactions, and S-adenosyl homocysteine (SAH). The crystal structure of the Mtb *sahH* (PDB: 2ziz, 2zj0, 2zj1, 3ce6; [40••]) reveals a similar architecture to *sahH* in other organisms. However, a co-crystal complex with SAH shows a distinct rotation of the imidazole ring of His363, which serves to open a solvent-access channel where density for this free portion of the substrate might bind (Figure 1). This contrasts with all previous liganded structures, which have only been solved in complex with adenosine or adenosine analogs, and in which room for the methionine appendage of the full substrate at the 5′ ribose position has been closed off.

htrA2 (Rv0983) is a heat-shock stress-response PDZ containing serine protease that gets upregulated under high-temperature conditions, similar to degP in *E. coli*. The structure of *htrA2* (PDB: 1y8t; [41]) reveals a hexameric structure, with the PDZ domains around the periphery, for recognition of C-terminal peptide signals, and serine-protease domains opening to an internal cavity (Figure 2). The enzyme also appears to have chaperone activity, as assays demonstrate that it can promote folding of various proteins, though the regulatory control of the various activities remains unclear.

fcoT (Rv0098) is a novel type III thioesterase that is thought to hydrolyze C16-C18 fatty acids from acyl-CoA. This enzyme has a 'hot-dog' fold (PDB: 2pfc; [42]) resembling type II thioesterases (beta-sheet curled around a long alpha-helix) but lacks the active site residues crucial to the catalytic mechanism in other organisms (Figure 3). The enzyme is part of an operon

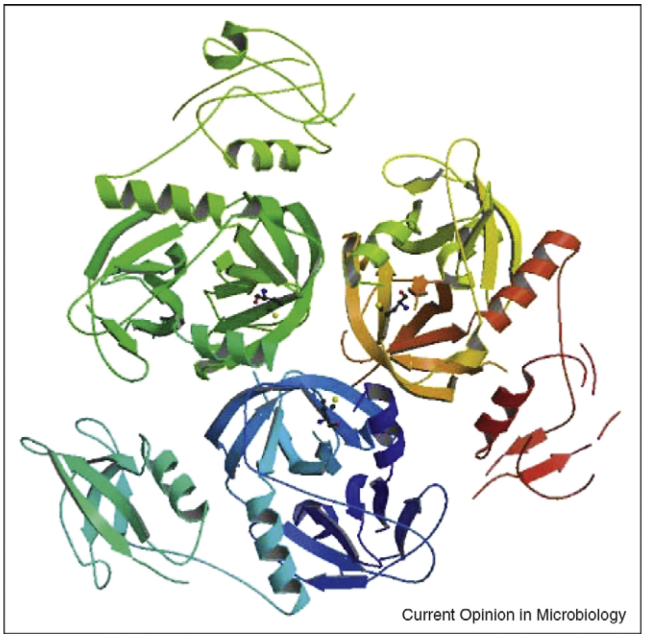
Figure 1



The active site of S-adenosyl homocysteine hydrolase (SahH; PDB 3dhy) in complex with a substrate analog (with a thioethyl group attached to NAD). The thioethyl appendage protrudes into a novel solvent access channel opened up by rotation upward of His363. The loop shown in cyan is part of a 37-residue insertion found in the *M. tuberculosis* form of the enzyme, along with other prokaryotes, but is not found in mammalian orthologs (adapted from [40••]).

(Rv0096-Rv0101) containing an acyl-CoA synthetase and a non-ribosomal peptide synthase that appears to define its function in producing a novel lipopeptide that, given the essentiality of *fcoT*, is also likely to be essential.

Figure 2



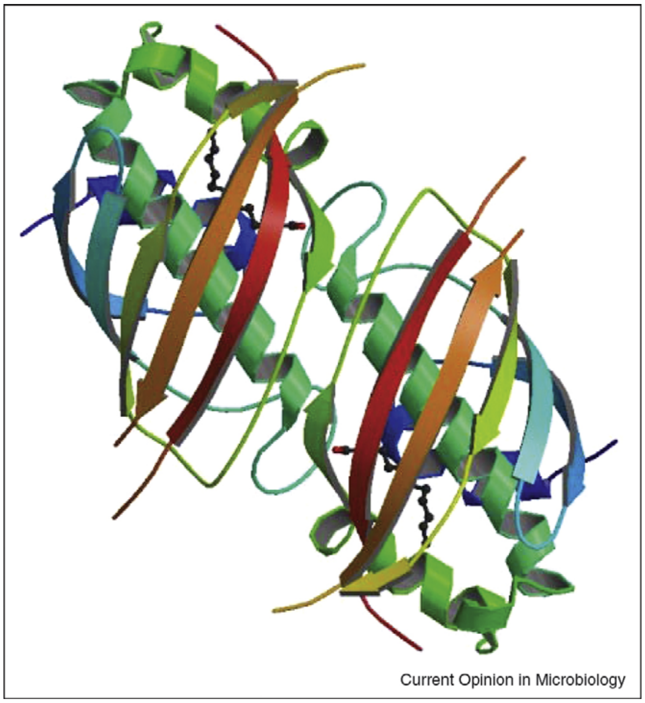
Backbone diagram of *htrA2* (PDB 1y8t [41]), shown as a homo-trimeric complex. The protease active sites are oriented toward the core of the complex (image generated on www.rcsb.org).

The TB genome contains only for a single annotated beta-lactamase gene (*blaC*, Rv2068c), which is thought to be responsible for conferring resistance to beta-lactams in mycobacteria, detoxifying these drugs by ring-opening before they can interfere with transpeptidases in cross-linking of peptidoglycan in the cell wall. The crystal structure (PDB: 2gdn; [43]) reveals amino acid substitutions in the active site, compared to other beta-lactamases in other organisms, that help explain the broad specificity against beta-lactams observed (at the apparent cost of lower overall activity). The structure might help in the development of more potent analogs of the covalent inhibitor clavulenate [44], which could be taken in combination with beta-lactams such as ipenem to potentiate them.

The structure of salicylate synthase, *mbtI (Rv2836c)*, has been solved (PDB: 2i6y, 2g5f; [45]). This enzyme converts chorismate to salicylate, as the first step toward biosynthesis of mycobactin, a siderophore used for scavenging iron from the extracellular environment. It has a structure similar to other chorismate-binding enzymes, such as TrpE and PabB.

Several proteins in the arginine biosynthesis pathway have been solved, including N-acetyl-gamma-glutamyl-phosphate reductase (*argC*, Rv1652, PDB: 2nqt, 2i3g, 2i3a; [46]), and ornithine transcarbamoylase (*argF*; Rv1656, PDB: 2i6u, 2pgp; [47]). All genes in this pathway have been found to be essential for growth *in vitro* using transposon-insertion [37]. *argC* is an NADP-dependent

Figure 3



Backbone diagram of the thioesterase *fcoT* (PDB 2pfc [42]), as a homodimer. A long-chain fatty acid is bound as a ligand to each subunit (image generated on www.rcsb.org).

enzyme that reductively dephoshoporylates N-acetyl-gamma-glutamyl-phosphate to yield a semi-aldehyde. Several small conformational changes were observed upon co-factor binding, relative to the apo form. *argF* converts ornithine into citrulline using carbamoyl phosphate. Binding of the latter produces striking loop movements of 7 and 12 Å for two loops covering the active site when the substrates are bound.

The VapBC-5 complex is the first toxin-antitoxin pair from TB to be solved (Rv0626 and Rv0627; PDB: 3dbo; [48••]). There are 38 pairs of toxin-antitoxin (TA) genes in the TB genome, including 23 in the VapBC family. The toxins in many TA pairs are ribonucleases that degrade mRNA in the absence of the antitoxin. VapC-5 is a 153-residue protein (PIN domain) with an alpha-beta core and a 2-helix 'clip' that protrudes from the core. Thirty-three of 86 residues of VapB-5 are observed to fold into an alternating sequence of helices and coil that lays in a groove between the core and the clip of VapC-5, burying the catalytic site. In *in vitro* assays, VapBC-5 shows weak RNase activity, though the substrate specificity, interaction with other cellular components, activation mechanism that leads to destabilization of the VapBC complex, and role of the subsequent RNA degradation in the mycobacterium life cycle remains unclear.

The structure of pantothenate synthetase (*panC*, Rv3602c) has been solved in both apo form and numerous complexes (PDB: 2a7x and others; [49,50]). *panC* is the second step in the biosynthesis pathway for pantothenate, a constituent of co-enzyme A as well as acyl-carrier protein (*acpM*, used in lipid biosynthesis). It condenses pantoate with beta-alanine, using ATP as a co-factor. The crystal structure of the protein reveals deep, distinct pockets for binding each of the three substrates. High-throughput screening has yielded a small-molecule inhibitor, nafronyl oxalate, with a $K_i \sim 73\ \mu M$ [51•], which has been observed crystallographically to bind in the ATP-binding subsite.

Future technology development: toward drug discovery

Structural genomics is one of a number of new technologies that, in combination, can lead toward drug discovery. One of the crucial steps in drug discovery is target identification and validation. Structure-based drug design requires knowledge of the proteins whose inhibition will be bacteriocidal for the mycobacteria. In some cases, these are on crucial metabolic pathways, though pathways are often redundant, and some pathways may become more or less sensitive under different conditions (e.g. such as the role of isocitrate lyase and malate synthase in supporting fatty-acid metabolism and growth on alternative carbon sources via the glyoxylate shunt; [52,53]). In other cases, the bacteriocidal effect may be due to other causes, such as build-up of toxic intermediates. One of the most recently developed approaches to evaluating gene essentiality is transposon-insertion [37].

There are various genetic tools as well for target validation, including construction of knockouts and knockdowns (conditional mutants whose expression can be turned off, for example, using a tetracycline-promoter; [54]). TB is difficult to work with genetically because it appears to lack a functional homologous recombination system. However, specialized phage-based transduction systems have been developed [55]. One approach especially useful for determining whether drugs are inhibiting their intended targets ('on-target') is overexpression [56].

Mycobacterium smegmatis is often used in the lab for preliminary target evaluation because it is non-virulent and shares much of the same genome, although *M. bovis* (and BCG) is closer evolutionarily, within the *M. tuberculosis* complex [57]. However, there are differences between *M. smegmatis* and *M. tuberculosis*, including different growth rates (*M. smegmatis* is fast-growing), different cell wall constituents, and so on, making it an imperfect model system. Recently, a significant advance has come from the development of a vaccine strain of *M. tuberculosis*, mc^2-7000, which is a pantothenate-auxotroph (knockout of *panC*) safe to work with under BL-2 conditions. This should allow greater fidelity in testing of drug cidality.

The TBSGC provides as a core service tools to help identify small-molecule inhibitors, either as drug leads or tool compounds. This can be automated through high-throughput screening as well as virtual screening. In the TBSGC a preliminary custom-designed 51 460-compound diversity library of compounds selected from a database of over three million commercially available, filtered by Lipinski's Rules for drug-likeness [58], and clustered to enforce non-redundancy. Using this library, and given an appropriate enzyme assay, a large number of compounds can be screened quickly to identify those with measurable inhibition activity (typically at around 50 μM concentrations). Also, fragment-screening techniques are being developed to identify binding of smaller compounds to different parts of a site (via methods such as isothermal titration calorimetry or NMR), which can sometimes be combined chemically to form larger molecules with near-additive gains in affinity (on a log scale; [59]).

In addition, computational techniques such as molecular docking can be used to perform virtual screening (in high-throughput fashion, distributed over a cluster of computers), which can be used to select compounds for testing that fit the target active site (both sterically and electrostatically), and has been shown to enrich for true competitive inhibitors [60]. For example, virtual screening of two databases totaling over 500 000 compounds was performed for inhibitors to ATP phosphoribosyltransferase (HisG, the first step in the histidine biosynthesis pathway) using both the GOLD and FlexX docking algorithms [61••]. This led to the discovery of several compounds with 4–6 μM enzyme inhibition activity (K_i), and one of these compounds had an MIC of 12.5 μM in a whole-cell assay against *M. smegmatis.*

Compounds discovered by high-throughput screening (using enzyme assays) and virtual screening are often limited, in that they may have only moderate inhibition activity (typically in the micro-molar IC_{50} range), or poor cell wall penetration (which can be evaluated with whole-cell assays), or be toxic to animals. In some cases, potency can be increased (or other pharmacokinetic properties adjusted) through medicinal chemistry, in the development of leads into viable drug candidates.

Nonetheless, the identification of small compounds with even moderate inhibition activity can be of great service as 'chemical tools' in probing the biological function of genes within the cell. Inhibitors can be used to selectively modulate that activity of a gene at specific times, and with variable degree, in a way that is much more precise and easy to control than with conditional mutant expression systems, which are frequently 'leaky' and which are sensitive to stability of previously synthesized mRNAs or polypeptides. Compound tools can even be used in combination with genetic knockdown experiments, where one reduces the level of a given protein target, typically with anti-sense RNA, and then uses the tool compound to reduce the remaining activity. This combination of methods often allows even micromolar inhibitors to have great utility in defining the importance of a novel target.

Finally, with the advent of new sequencing technologies, such as the Illumina Genome Analyzer (Illumina, Inc.), rapid and inexpensive whole-genome sequencing is now becoming a reality. These machines provide high coverage ($\sim$50×) with very short read lengths (36–72 bp), and can sequence an entire genome within $\sim$24 h. One way to apply this to target validation is to sequence strains that are resistant to a particular drug. Drug resistance (often occurring at a natural rate of around 10^{-8}) can arise from many sources [62]. The most straightforward are mutations in the active site of the target itself, although mutations leading to resistance can also occur in upstream-regulatory regions (causing upregulation or downregulation), in pro-drug activators, in drug efflux pumps, or in detoxifying enzymes such as acetylases, esterases, and so on. In cases where the target of a drug is not known, whole-genome sequencing, combined with SNP analysis against other known drug-sensitive or drug-resistant strains, can help narrow down the target. In other cases, whole-genome sequencing of clinical isolates or isogenic laboratory mutants that are known not to carry mutations in expected positions can help understand the mechanism of action by identifying other proteins that may be involved (as in the connection between mutations in mycothiol biosynthesis enzymes and ethionamide resistance; [63]). We have implemented this as a core function in the TBSGC.

When combined together within an integrated structural genomics pipeline, these technologies can provide powerful tools for drug discovery against pathogenic organisms, such as drug-resistant tuberculosis, by coupling structure determination with high-throughput screening, structure-based drug design, genetic methods for target validation, and whole-genome sequencing.

Acknowledgements

This work was supported partly by NIH grant PO1AI068135 and the Robert A Welch Foundation (JCS). The authors would also like to thank the members of the Tuberculosis Structural Genomics Consortium.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hendrickson W, Ogata C: **Phase determination from multiwavelength anomalous diffraction measurements**. *Methods Enzymol* 1997, **276**:494-523.

2. Dauter Z: **New approaches to high-throughput phasing**. *Curr Opin Struct Biol* 2002, **12**:674-678.

3. Adams PD, Pannu NS, Read RJ, Brünger AT: **Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement**. *Proc Natl Acad Sci U S A* 1997, **94(May (10))**:5018-5023.

4. Terwilliger TC, Berendzen J: **Automated MAD and MIR structure solution**. *Acta Crystallogr* 1999, **D55**:849-861.

5. Morris RJ, Perrakis A, Lamzin VS: **ARP/wARP and automatic interpretation of protein electron density maps**. *Methods Enzymol* 2003, **374**:229-244.

6. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ: **Phaser crystallographic software**. *J Appl Cryst* 2007, **40**:658-674.

7. • Manjasetty BA, Turnbull AP, Panjikar S, Büssow K, Chance MR: **Automated technologies and novel techniques to accelerate protein crystallography for structural genomics**. *Proteomics* 2008, **8(February (4))**:612-625.

A review of high-throughput methods for structural genomics, ranging from protein production and crystallization, to data collection and processing.

8. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S: **Structural genomics: beyond the human genome project**. *Nature Genet* 1999, **23**:151-157.

9. Chothia C: **Proteins. One thousand families for the molecular biologist**. *Nature* 1992, **357(June (6379))**:543-544.

10. Grant A, Lee D, Orengo C: **Progress towards mapping the universe of protein folds**. *Genome Biol* 2004, **5(5)**:107.

11. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B: **Automatic target selection for structural genomics on eukaryotes**. *Proteins* 2004, **56(August (2))**:188-200.

12. Finn RD, Tate J, Mistry J, Coggill PC, Sammut JS, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database**. *Nucleic Acids Res* 2008, **36**:D281-D288.

13. Kelley LA, Sternberg MJE: **Protein structure prediction on the web: a case study using the Phyre server**. *Nat Protoc* 2009, **4**:363-371.

14. Schmid MB: **Seeing is believing: the impact of structural genomics on antimicrobial drug discovery**. *Nat Rev Microbiol* 2004, **2**:739-746.

15. Weigelt J, McBroom-Cerajewski LD, Schapira M, Zhao Y, Arrowmsmith CH: **Structural genomics and drug discovery: all in the family**. *Curr Opin Chem Biol* 2008, **12(February (1))**:32-39.

16. Congreve M, Murray CW, Blundell TL: **Structural biology and drug discovery**. *Drug Discov Today* 2005, **10(July (13))**:895-907.

17. Barker JJ: **Antibacterial drug discovery and structure-based design**. *Drug Discov Today* 2006, **11(May (9–10))**:391-404.

18. •• Nair R, Liu J, Soong TT, Acton TB, Everett JK, Kouranov A, Fiser A, Godzik A, Jaroszewski L, Orengo C, Montelione GT, Rost B: **Structural genomics is the largest contributor of novel structural leverage**. *J Struct Funct Genomics* 2009, **1(April (2))**:181-191.

Up-to-date statistics on protein structures contributed to the RCSB by Structural Genomics Centers.

19. •• Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D: **Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis***. *Proc Natl Acad Sci U S A* 2006, **103(May (21))**:8060-8065.

Describes a large-scale screening effort to identify complexes between PPE and PGRS proteins that could be crystallized and solved. This is the first crystal structure from each of these highly duplicated families in the TB genome, whose function remains obscure.

20. Columbus L, Lipfert J, Klock H, Millett I, Doniach S, Lesley SA: **Expression, purification, and characterization of *Thermotoga maritima* membrane proteins for structure determination**. *Protein Sci* 2006, **15(May (5))**:961-975.

21. Arcus VL, Lott JS, Johnston JM, Baker EN: **The potential impact of structural genomics on *Mycobacterium tuberculosis* drug discovery**. *Drug Discov Today* 2006, **11**:28-34.

22. • Murillo AC, Li HY, Alber T, Baker EN, Berger JM, Cherney LT, Cherney MM, Cho YS, Eisenberg D, Garen CR *et al.*: **High throughput crystallography of TB drug targets**. *Infect Disord Drug Targets* 2007, **7(June (2))**:127-139.

This review contains additional details about the high-throughput technologies used in the Tuberculosis Structural Genomics Consortium.

23. World Health Organization: Global tuberculosis control—epidemiology, strategy, financing. http://www.who.int/tb/publications/global_report/2009/en/index.html, accessed on 3/29/2009.

24. •• Jassal M, Bishai WR: **Extensively drug-resistant tuberculosis**. *Lancet Infect Dis* 2009, **9(January (1))**:19-30.

A recent review of XDR-TB, including world-wide epidemiology, genome sequence analysis of drug-resistance mutations, and clinical aspects.

25. Gomez JE, McKinney JD: ***M. tuberculosis* persistence, latency, and drug tolerance**. *Tuberculosis (Edinb)* 2004, **84(1–2)**:29-44.

26. Russell DG: **Who puts the tubercle in tuberculosis?** *Nat Rev Microbiol* 2007, **5(January (1))**:39-47.

27. Johnston JM, Arcus VL, Morton CJ, Parker MW, Baker EN: **Crystal structure of a putative methyltransferase from *Mycobacterium tuberculosis*: misannotation of a genome clarified by protein structural analysis**. *J Bacteriol* 2003, **185(July (14))**:4057-4065.

28. Lee K, Zhan X, Gao J, Qiu J, Feng Y, Meganathan R, Cohen SN, Georgiou G: **RraA: a protein inhibitor of RNase E activity that globally modulates RNA abundance in *E. coli***. *Cell* 2003, **114(September (5))**:623-634.

29. • Goldstone RM, Moreland NJ, Bashiri G, Baker EN, Lott JS: **A new Gateway® vector and expression protocol for fast and efficient recombinant protein expression in *Mycobacterium smegmatis***. *Protein Expr Purif* 2008, **57**:81-87.

Describes an efficient system for cloning genes in preparation for subsequent expression and crystallization. All ~4000 genes in the tuberculosis genome have been cloned into Gateway vectors.

30. Bursey EH, Kim C-Y, Yu M, Terwilliger TC, Hung L-W: **An automated high-throughput screening method for the identification of high-yield, soluble protein variants using cell-free expression and systematic truncation**. *J Struct Funct Genomics* 2006, **7(3–4)**:139-147.

31. Yang JK, Park MS, Waldo GS, Suh SW: **Directed evolution approach to a structural genomics project: Rv2002 from *Mycobacterium tuberculosis***. *Proc Natl Acad Sci U S A* 2003, **100(January (2))**:455-460.

32. • Goldschmidt L, Cooper D, Derewenda Z, Eisenberg D: **Toward rational protein crystallization: a Web server for the design of crystallizable protein variants**. *Protein Sci* 2007, **16**:1569-1576.

Describes the method of Surface Entropy Reduction for identifying mutations to make in proteins to facilitate crystallization.

33. Rupp B, Segelke BW, Krupka HI, Lekin T, Schäfer J, Zemla A, Toppani D, Snell G, Earnest T: **The TB structural genomics consortium crystallization facility: towards automation from protein to electron density**. *Acta Crystallogr* 2002, **D58**:1514-1518.

34. Hansen CL, Skordalakes E, Berger JM, Quake SR: **A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion**. *Proc Natl Acad Sci U S A* 2002, **99**:16531-16536.

35. Adams PD, Grosse-Kunstleve RW, Hung L-W, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC: **PHENIX: building new software for automated crystallographic structure determination**. *Acta Crystallogr* 2002, **D58**:1948-1954.

36. Hasan S, Daugelat S, Rao PS, Schreiber M: **Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis***. *PLoS Comput Biol* 2006, **2(June (6))**:e61.

37. Sassetti CM, Boyd DH, Rubin EJ: **Comprehensive identification of conditionally essential genes in mycobacteria**. *Proc Natl Acad Sci U S A* 2001, **98(June (22))**:12712-12717.

38. Smith DA, Parish T, Stoker NG, Bancroft GJ: **Characterization of auxotrophic mutants of *Mycobacterium tuberculosis* and their potential as vaccine candidates**. *Infect Immun* 2001, **69**:1142-1150.

39. Lee CE, Goodfellow C, Javid-Majd F, Baker EN, Shaun Lott J: **The crystal structure of TrpD, a metabolic enzyme essential for lung colonization by *Mycobacterium tuberculosis*, in complex with its substrate phosphoribosylpyrophosphate**. *J Mol Biol* 2006, **355(4)**:784-797.

40. •• Reddy MC, Kuppan G, Shetty ND, Owen JL, Ioerger TR, Sacchettini JC: **Crystal structures of *Mycobacterium tuberculosis* S-adenosyl-L-homocysteine hydrolase in ternary complex with substrate and inhibitors**. *Protein Sci* 2008, **17(12)**:2134-2144.

Describes the structure of *M. tuberculosis* SahH, which is an essential protein involved in re-cycling co-factor SAM used in methylation. Crystal complexes with several adenosine analogs are also reported for this potential drug target.

41. Mohamedmohaideen NN, Palaninathan SK, Morin PM, Williams BJ, Braunstein M, Tichy SE, Locker J, Russell DH, Jacobs WR Jr, Sacchettini JCL: **Structure and function of the virulence-associated high-temperature requirement A of *Mycobacterium tuberculosis***. *Biochemistry* 2008, **47(23)**:6092-6102.

42. Wang F, Langley R, Gulten G, Wang L, Sacchettini JC: **Identification of a type III thioesterase reveals the function of an operon crucial for Mtb virulence**. *Chem Biol* 2007, **14(5)**:543-551.

43. Wang F, Cassidy C, Sacchettini JC: **Crystal structure and activity studies of the *Mycobacterium tuberculosis* beta-lactamase reveal its critical role in resistance to beta-lactam antibiotics**. *Antimicrob Agents Chemother* 2006, **50(8)**: 2762-2771.

44. Hugonnet JE, Blanchard JS: **Irreversible inhibition of the *Mycobacterium tuberculosis* beta-lactamase by clavulanate**. *Biochemistry* 2007, **46(October (43))**:11998-12004.

45. Harrison AJ, Yu M, Gårdenborg T, Middleditch M, Ramsay RJ, Baker EN, Lott JS: **The structure of MbtI from *Mycobacterium tuberculosis*, the first enzyme in the biosynthesis of the siderophore mycobactin, reveals it to be a salicylate synthase**. *J Bacteriol* 2006, **188(September (17))**:6081-6091.

46. Cherney LT, Cherney MM, Garen CR, Niu C, Moradian F, James MN: **Crystal structure of N-acetyl-gamma-glutamyl-phosphate reductase from *Mycobacterium tuberculosis* in complex with NADP(+)**. *J Mol Biol* 2007, **367(April (5))**: 1357-1369.

47. Sankaranarayanan R, Cherney MM, Cherney LT, Garen CR, Moradian F, James MN: **The crystal structures of ornithine carbamoyltransferase from *Mycobacterium tuberculosis* and its ternary complex with carbamoyl phosphate and L-norvaline reveal the enzyme's catalytic mechanism**. *J Mol Biol* 2008, **375(January (4))**:1052-1063.

48. •• Miallau L, Faller M, Chiang J, Arbing M, Guo F, Cascio D, Eisenberg D: **Structure and proposed activity of a member of the VapBC family of toxin-antitoxin systems: VapBC-5 from *Mycobacterium tuberculosis***. *J Biol Chem* 2009, **284(January (1))**:276-283.

The papers describes the first crystal structure of the complex of a toxin-antitoxin pair in the *M. tuberculosis* genome, in which there are 38 such pairs. Analysis of the structure of the toxin suggests that in may have a ribonuclease function.

49. Wang S, Eisenberg D: **Crystal structures of a pantothenate synthetase from *M. tuberculosis* and its complexes with substrates and a reaction intermediate**. *Protein Sci* 2003, **12(May (5))**:1097-1108.

50. Wang S, Eisenberg D: **Crystal structure of the pantothenate synthetase from *Mycobacterium tuberculosis*, snapshots of the enzyme in action**. *Biochemistry* 2006, **45**:1554-1561.

51. • White EL, Southworth K, Ross L, Cooley S, Gill RB, Sosa MI, Manouvakhova A, Rasmussen L, Goulding C, Eisenberg D, Fletcher TM 3rd: **A novel inhibitor of *Mycobacterium tuberculosis* pantothenate synthetase**. *J Biomol Screen* 2007, **12(February (1))**:100-105.

This paper reports the discovery of nafronyl oxalate as an inhibitor of *M. tuberculosis* PanC, along with a crystal structure of the complex showing how it binds in the active site.

52. McKinney JD, Höner zu Bentrup K, Muñoz-Elías EJ, Miczak A, Chen B, Chan WT, Swenson D, Sacchettini JC, Jacobs WR Jr, Russell DG: **Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase**. *Nature* 2000, **406(August (6797))**:735-738.

53. Smith CV, Huang CC, Miczak A, Russell DG, Sacchettini JC, Höner zu Bentrup K: **Biochemical and structural studies of malate synthase from *Mycobacterium tuberculosis***. *J Biol Chem* 2003, **278(January (3))**:1735-1743.

54. Carroll P, Muttucumaru DG, Parish T: **Use of a tetracycline-inducible system for conditional expression in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis***. *Appl Environ Microbiol* 2005, **71(June (6))**:3077-3084.

55. Bardarov S Jr, Pavelka MS Jr, Sambandamurthy V, Larsen M, Tufariello J, Chan J, Hatfull G, Jacobs WR Jr: **Specialized transduction: an efficient method for generating marked and unmarked targeted gene disruptions in *Mycobacterium tuberculosis*, *M. bovis* BCG, and *M. smegmatis***. *Microbiology* 2002, **148**:3007-3017.

56. Larsen MH, Vilchèze C, Kremer L, Besra GS, Parsons L, Salfinger M, Heifets L, Hazbon MH, Alland D, Sacchettini JC, Jacobs WR Jr: **Overexpression of inhA, but not kasA, confers resistance to isoniazid and ethionamide in *Mycobacterium smegmatis*, *M. bovis* BCG and *M. tuberculosis***. *Mol Microbiol* 2002, **46(October (2))**:453-466.

57. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K *et al.*: **A new evolutionary scenario for the *Mycobacterium tuberculosis* complex**. *Proc Natl Acad Sci U S A* 2002, **99**:3684-3689.

58. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings**. *Adv Drug Deliv Rev* 2001, **46**:3-26.

59. Congreve M, Chessari G, Tisi D, Woodhead AJ: **Recent developments in fragment-based drug discovery**. *J Med Chem* 2008, **51(13)**:3661-3680.

60. Klebe G: **Virtual ligand screening: strategies, perspectives and limitations**. *Drug Discov Today* 2006, **11**:580-594.

61. •• Cho Y, Ioerger TR, Sacchettini JC: **Discovery of novel nitrobenzothiazole inhibitors for *Mycobacterium tuberculosis* ATP phosphoribosyl transferase (HisG) through virtual screening**. *J Med Chem* 2008, **51(19)**:5984-5992.

Describes the successful application of virtual screening to identify compounds that inhibit *M. tuberculosis* HisG based on a crystal structure of the complex with AMP and histidine.

62. Gillespie SH: **Evolution of drug resistance in *Mycobacterium tuberculosis*: clinical and molecular perspective**. *Antimicrob Agents Chemother* 2002, **46(February (2))**:267-274.

63. Vilchèze C, Av-Gay Y, Attarian R, Liu Z, Hazbón MH, Colangeli R, Chen B, Liu W, Alland D, Sacchettini JC, Jacobs WR Jr: **Mycothiol biosynthesis is essential for ethionamide susceptibility in *Mycobacterium tuberculosis***. *Mol Microbiol* 2008, **69(September (5))**:1316-1329.