

Structure-guided, target-based drug discovery – exploiting genome information from HIV to mycobacterial infections

Sony Malhotra

Sherine E. Thomas

Bernardo Ochoa Montano

Tom L. Blundell✉

Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, UK

✉Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Old Addenbrooke's Site, Cambridge CB2 1GA, UK; e-mail: tlb20@hermes.cam.ac.uk or tom@cryst.bioc.cam.ac.uk

Received: May 31, 2016

Accepted: June 14, 2016

Key words: structure-guided, fragment-based, drug design/discovery, HIV protease, inhibitors, *Mycobacterium tuberculosis*, *Mycobacterium abscessus*

Abbreviations: IGF – Insulin-like growth factor; FBDD – Fragment-based drug design; TB – Tuberculosis; MTb – *Mycobacterium tuberculosis*; MAb – *Mycobacterium abscessus*; HIV – Human Immunodeficiency Virus; AIDS – Acquired Immune Deficiency Syndrome

Acknowledgements: SM is funded by a joint grant between the Medical Research Council, UK and Department of Biotechnology (India); SET is funded by the Cystic Fibrosis Trust; BOM is funded by the Bill and Melinda Gates Foundation; TLB is funded by the Wellcome Trust.

ABSTRACT

The use of protein crystallography in structure-guided drug discovery allows identification of potential inhibitor-binding sites and optimisation of interactions of hits and lead compounds with a target protein. An early example of this approach was the use of the structure of HIV protease in designing AIDS antivirals. More recently, use of structure-guided design with fragment-based drug discovery, which reduces the size of screening libraries by decreasing complexity, has improved ligand efficiency in drug design. Here, we discuss the use of structure-guided target identification and lead optimisation using fragment-based approaches in the development of new antimicrobials for mycobacterial infections.

INTRODUCTION

In the 1960s and 70s, when protein crystallography was still in its infancy, there was an emerging awareness of its potential contribution to making new medicines. Mutations resulting in abnormal haemoglobins were recognised as the genetic defects that resulted in inherited single-gene disorders, such as sickle-cell disease. Max Perutz and his team in Cambridge discussed the structural impacts of these mutations on oxygen affinity and subunit cooperativity. Dorothy Hodgkin's Oxford laboratory members often visited Eli Lilly, Wellcome Foundation and Novo, where crystalline insulin was produced for treatment of diabetes. Insulin biochemistry and crystallisation had been discussed by Jørgen Schlichtkrull of Novo [1] and the design of slow acting insulins had long been contemplated. In the late 60s, when the structure of insulin was defined [2,3], the availability of many amino-acid sequences from Fred Sanger's laboratory in Cambridge stimulated ideas about insulin receptor binding and discussion about new insulin designs in the Hodgkin laboratory in Oxford. During the same period, the Aachen, New York and Shanghai groups completed the synthesis of insulin, allowing ideas about novel synthetic insulins to be seriously considered.

As early as the 1960s, David Phillips and his team had begun to think about building 3D models of homologues that may be of biological or biomedical relevance, the first being α -lactalbumin based on lysozyme using wire models [4]. In the 1970s, homology or comparative modelling of clinically more relevant targets, such as relaxin and somatomedins C (IGF) based on insulin structures [5-8], was made more realistic by the availability of graphics facilities and software for manipulating them, most importantly Frodo from Alwyn Jones [9]. Eventually, clinically important drug targets like renin, the aspartic protease that cleaves angiotensinogen and converts it into angiotensin I, an essential step in regulating blood pressure, were modelled on less exciting enzymes such as fungal pepsins, allowing structural biology to move into big pharma [10]. Indeed, many companies acquired models of renin from the Blundell lab and computational structure-guided drug discovery became a focus in the pharma industry in the 1980s.

However, at the same time structures were also defined experimentally of apo-enzymes and complexes of renin and its close homologues [10-15]. For other molecules, such as nerve growth factor, suggestions of homology with insulin, arising from the existence of three disulphide bonds in both proteins, were not corroborated by sequence analysis [16]. Blundell in the 70s moved to start crystallography on nerve growth factor, only to discover that Alex Wlodawer had been working on the crystal structure with Eric Shooter and Keith Hodgson, since the early 1970s [17]. At that early stage they agreed to collaborate on this project and publish together. Although it took a while, they did so in 1991 [18].

STRUCTURE OF HIV PROTEASE AND STRUCTURE-BASED DRUG DESIGN

In the 80s many groups, including those of Wlodawer and Blundell, worked in parallel on the structural investigations of aspartic proteases (then known as acid proteases), focusing increasingly on retroviral enzymes.

In 1978, Blundell along with Jordan Tang, who had determined the sequence of pepsin, and Mike James, who was also defining X-ray structures of aspartic proteases, published a hypothesis that these proteases had evolved from an ancestral dimer by gene duplication, fusion and divergence to give more active enzymes [19]. The questions this posed were: "Does a similar dimeric enzyme exist in nature today? What does this enzyme do?" The answer came in the retroviral proteases, most topically in HIV, the genome of which encodes polyproteins, which are cleaved by this dimeric viral protease to produce the mature viral proteins that are essential to generate infectious viral particles (Fig. 1). The Blundell lab identified the un-annotated enzyme protomer sequence in the genome but it was first published by Toh *et al.* in 1985 [20]. A model by Pearl and

Taylor, then based in Birkbeck, was published in 1987 [21]. In early 1989, Wlodawer and colleagues published the first structure of a retroviral protease, that of Rous Sarcoma Virus [22]. Almost simultaneously, Merck published a structure of HIV protease [23,24]; this structure was updated by two experimental structures determined independently by the Blundell [25] and Wlodawer [22] labs. Recently, a structure from Wlodawer lab of two *Rickettsia conorii* constructs was determined, with structurally similar folds to that of the HIV aspartic protease, but corresponding to only half of the homodimer of HIV aspartic protease. This might be an ancestor of the monomeric and dimeric aspartic proteases [26]. Wlodawer's lab has also worked on the aspartic proteases from a variety of species such as yeast [27,28], plants [29,30], and other viruses [31-34].

LESSONS LEARNT FROM HIV PROTEASE ANALYSES

In summary, analyses of the HIV genome and proteome led to the discovery of a new protease that was critical to the discovery of new AIDS antivirals. It soon became evident that inhibiting the viral protease would block the viral replication and the production of mature infectious viral particles, thereby, disrupting the ability to infect new host cells (Fig. 1). The subsequent development of new AIDS antivirals demonstrated the importance of understanding the genome not only in terms of functions of gene products, but also their architectures for use in structure-guided drug discovery.

The homology between the renins and retroviral proteases led to suggestions of molecules to start new campaigns of drug discovery and these provided important molecules used in discovering new therapeutics against AIDS. However, a more general approach to screening potential targets and exploring chemical space - was clearly required if the knowledge of the architectures of gene products was to be fully exploited in drug discovery. Over the subsequent 25 years, several new approaches have been introduced that exploit knowledge of the architecture of the targets. Two of the most influential have been the computational identification of putative interacting compounds with the target protein using virtual screening and experimental structure-guided fragment-based drug discovery.

VIRTUAL SCREENING AND COMPUTATIONAL PREDICTION OF BINDING HOTSPOTS

Virtual screening allows a large number of small molecules to be screened *in silico* for their binding sites [35]. Several computational programs such as GRID [36], LUDI [37], Sprout [38] and MUSIC [39] are available for positioning these small molecules in binding hotspots. These methods explore electrostatic, van der Waals or hydrogen-bonding interactions involved in the molecular recognition [40]. Multiple Copy Simultaneous Search (MCSS) determines energetically

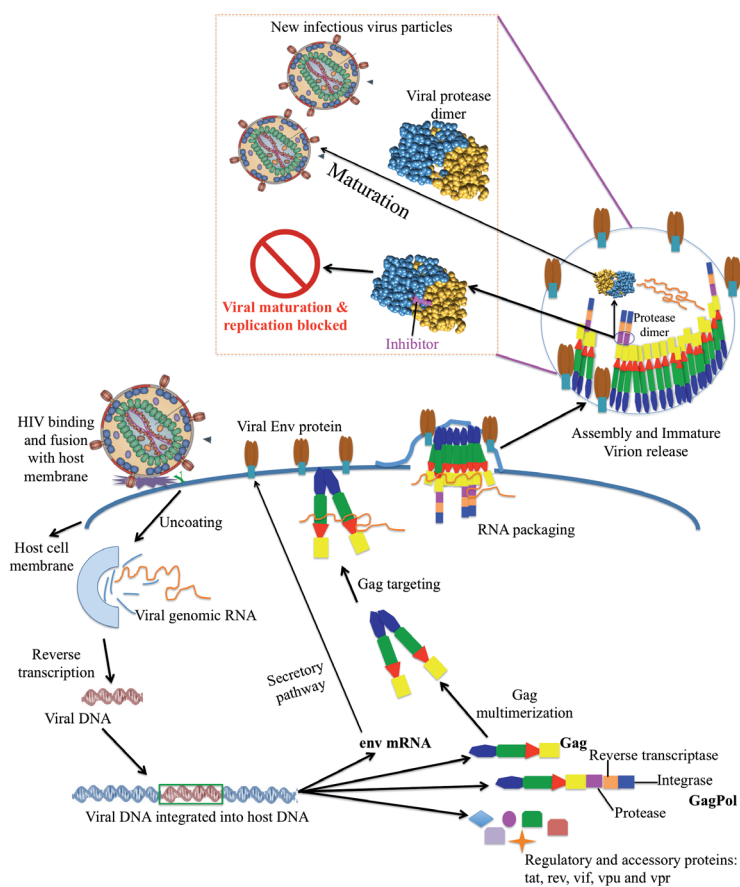


Figure 1. Design of anti-HIV inhibitors by targeting viral protease. Once the viral DNA is translated to Gag and GagPol polyproteins, Gag multimerises and is targeted to the host cell membrane where it anchors itself to the lipid rafts. This initiates the assembly of viral particles at the membrane, which incorporates viral RNA and Env protein that reaches cell membrane *via* secretory pathway. This non-infectious virion then buds off from the membrane and maturation of viral particle is triggered by the viral protease, which is active in the dimeric form of the polyprotein and which cleaves Gag and GagPol polyproteins, producing proteins that are required for the formation of the capsid core and infectious viral particle. Designing inhibitors (shown in magenta) that bind the viral protease will block the cleavage of viral structural proteins from the polyproteins and hence inhibit the viral replication and formation of new infectious viral particles. From [120], modified.

favourable orientations and positions of functional groups to map the binding site [41]. Another commonly used program, SUPERSTAR from the Cambridge Crystallographic Data Centre (CCDC), utilizes a knowledge-based approach where intermolecular interactions found in high-resolution X-ray crystal structures of small molecules are used to calculate the binding propensities of functional groups in a protein hotspot [42,43]. Our group has recently introduced a new method using SUPERSTAR for finding atomic hotspots in the binding pocket of the target protein, by exploring the presence of donor, acceptor and hydrophobic probes and weighting them by the depth of the pocket. Subsequently, these are sampled using different molecular probes for their binding and the binding atoms are mapped back onto the structure resulting in a fragment-hotspot map [44].

Virtual screening methods (such as GOLD [45], AUTODOCK [46], GLIDE [47,48]) can be used to dock a diverse set of drug-like compounds to a target protein in a short time with nearly 50% of the molecules docked within 2 Å RMSD (root mean square deviation) of the experimentally determined structures [35]. Upon docking, these interactions are scored in order to discriminate positive hits from a large library of compounds and hence identify true interactions. The scoring functions include physical terms, such as hydrophobicity, solvation electrostatics, hydrogen bonding, ligand deformation energy and van der Waals interaction energy. However, while evaluating virtual screening results, one needs to be aware of the caveats and limitations, such as scoring the docked poses, especially when ligands/fragments bind with low affinity, and when there is protein flexibility and well organised solvation of the protein [40,49].

STRUCTURE-GUIDED FRAGMENT-BASED DRUG DISCOVERY

Fragment-based drug design (FBDD) has seen applications in lead identification for targets in several pharmaceutical companies. In FBDD, a collection of small mole-

cules (<300 Da), known as a fragment library, are screened against a target of interest, resulting in identification of initial hits. These are then optimized into lead candidates by chemically growing or linking the fragments, thereby exploring the chemical space available for binding to the target protein very effectively. A high-affinity lead molecule thus developed from a fragment hit retains the key binding interactions of the original fragment. Although low-molecular-weight fragments have relatively lower potency than the more complex molecules found in typical high-throughput screening (HTS) compound libraries, small fragments that bind, do so by making well defined and directional high-quality interactions and by displacing unhappy water molecules at hotspots on the protein. Compounds optimized from the resulting fragment hits are capable of achieving higher binding efficiency per atom and better physicochemical properties in comparison to those from HTS approach [50]. The principal advantage of the fragment-based approach over conventional drug discovery campaign is the requirement of relatively smaller libraries (typically hundreds or at most a few thousand fragments) of low-molecular-weight compounds to obtain good starting hits for lead discovery, in comparison to the larger libraries of $\sim 10^6$ or more compounds required in the HTS methods [51].

We are exploring the use of structure-guided FBDD for developing antimicrobials targeting proteins from Mycobacterial species. We use a combination of biophysical, and biochemical techniques in addition to X-ray crystallography to monitor hit identification, validation and subsequent elaboration into lead molecules, as summarised in figure 2.

In some campaigns where an apo-protein crystal structure is not available, the pipeline can be complemented by substructure searches on commercially available compound databases and/or *in-silico* screening predictions of analogues [35,52]. Winter *et al.* have demonstrated a successful application of fragment screening, in the absence of a suitable crystal structure, to identify hits that disrupt the interaction between NK1 and Met

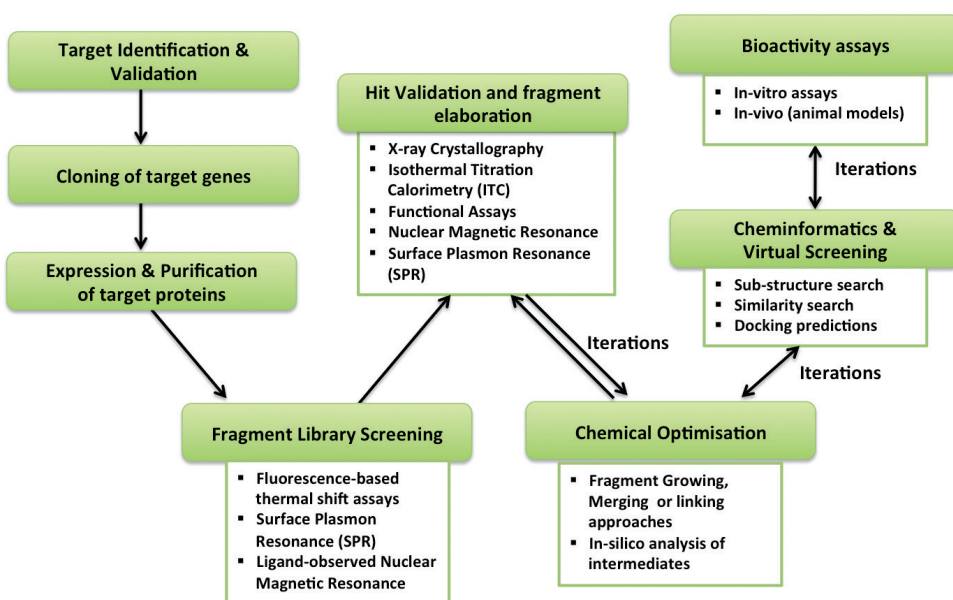


Figure 2. Schematic of the use of fragment-based drug discovery for identifying fragments that bind at hotspots on the target protein allowing subsequent elaboration into lead molecules.

and subsequently inhibit Met signalling [53]. Sub-structure searches on initial hits identified by biophysical methods such as surface plasmon resonance, were followed by docking predictions to identify compounds, with significant anti-tumorigenic and anti-migratory activity in cell-based assays.

CHALLENGES OF LARGER GENOMES – AN EXAMPLE OF MYCOBACTERIUM

We now illustrate how a combination of efficient exploration of biological space and use of fragment-based approaches and virtual screening to explore chemical space can give useful leads in drug discovery for my-

cobacteria with larger genomes encoding four to six thousand genes.

The family *Mycobacteriaceae* with 174 known species [54] is a part of one of the largest groups (Actinobacteria) of bacteria. It constitutes the acid-fast bacteria with high GC content and much thicker hydrophobic outer cell wall (containing mycolic acids) than many other bacteria. The family includes several pathogenic bacteria such as *Mycobacterium tuberculosis* (MTb, the causative agent of tuberculosis), *Mycobacterium leprae* (MLep, causes leprosy), *Mycobacterium abscessus* (MAb, rapid-growing opportunistic pathogen that causes chronic lung diseases) and free-living non-pathogenic bacteria such as *Mycobacterium kansasii*, *Mycobacterium vanbaalenii*, and *Mycobacterium smegmatis*. According to the most recent WHO report, 9.6 million people were diagnosed with tuberculosis (TB) and 1.5 million died because of the infection [55]. In 2015, one-in-three deaths of HIV-positive patients were due to tuberculosis [55]. MAb is a nontuberculous bacterium commonly found in water and soil and is a distant relative of MTb and MLep [56,57]. Unlike MTb, it is a rapidly growing organism that causes skin and soft tissues infections and pulmonary diseases, particularly in patients with lung disease (such as cystic fibrosis) or previous history of tuberculosis [57,58].

With the advances in the genomics technology over the past decade, there has been an increase in availability of whole genome sequences. This paves the way for analyses of whole genomes and cross-genome comparisons, which will help in understanding the genome organisation and the mechanisms leading to bacterial pathogenesis. 456 genomes are available in NCBI Genomes for different species and strains of mycobacteria [59] (April 2016). These sequences allow genome-scale studies and annotation of the genomes for structure and function.

GENOMES OF MYCOBACTERIUM

Before the genome sequence was made available for MTb, Sreevatsan *et al.* had analysed a 2Mb region of 26 structural genes in different strains and observed that the MTb genome has restricted allelic variation and lacks silent mutations in the analysed regions [60]. Studies of this type gave hints about the usefulness of the whole-genome sequences of these bacterial pathogenic organisms. The genome of the slow growing microorganism - MTb (strain H37Rv) was first sequenced in 1998 [61], paving the way for further investigation of specific gene families as well as structure and function annotation of the gene products. H37Rv has a large GC-rich genome (65.6%) of 4.4 Mb in size which codes for ~4000 genes [61], 6% of which are involved in lipid metabolism [61,62].

MAb (strain ATCC19977) was first isolated in 1953 from a knee abscess of a woman patient by Moore and Frerichs [63], but it was only in 2009 that the genome was sequenced and made available [64]. MAb has a larger genome than MTb (~5 Mb) and possesses ~5000 predicted coding sequences. Several gene clusters in synteny with non-mycobacterial species (such as *Nocardia sp.*, *Rhodococcus sp.*, *Pseudomonas aeruginosa*, *Burkholderia cepacia*) were observed implying that these genes might have been ac-

quired by horizontal gene transfer [64]; this is not unexpected given MAb's co-presence with other pulmonary infectious bacteria.

MAb genome encodes ~1000 more proteins than MTb. We obtained the protein sequences for MTb (strain H37Rv) and MAb (strain ATCC19977) genomes from UniProt [65] and compared them for the presence of orthologs using ProteinOrtho [66]. 2277 orthologous clusters were observed, covering 2308 proteins from MTb and 2333 MAb proteins as either orthologs or in-paralogs. All of these clusters appear to be of reasonable quality as indicated by the statistical measures obtained upon BLAST [67] comparisons. To exemplify the functions and relevance of these clusters, we are discussing three clusters, each of which comprises five either secretory components or proteins involved in secretory mechanisms (Tab. 1), suggesting the importance of these functions (and their possible role in virulence) in the two Mycobacterial species. The first cluster consists of two MTb proteins - EsxG and EsxS, which are part of Esx secretion system, and three MAb proteins annotated as probable PE family proteins. It has been proposed recently that the EsxGH heterodimer is secreted by MTb and is responsible for iron-uptake capacity inside the host and also plays an iron-independent role in host virulence [68]. Recently, the structure of EsxRS, a homolog of EsxGH, has been observed to form higher order oligomers by domain-swapping, which could be a mechanism employed by MTb to increase the strength of interaction of Esx members with the host receptors [69]. The second cluster is populated with MmpL (mycobacterial membrane protein large) protein0s, which are efflux pumps for a variety of outer membrane lipids and mycolic acids [70]. MmpL proteins are present in both slow and fast growing species of mycobacteria [71] and they may be associated with drug resistance in both MTb and nontuberculous bacteria by transporting antibiotics out of the periplasm [71-74]. The third cluster comprises two proteins of MTb (Diacylglycerol acyltransferase/mycolyltransferase Ag85C) and three MAb proteins annotated, as Antigen 85-A/B/C. Antigen 85 complex is known to be a secreted product in MTb with the ability to bind fibronectin, which in turn contributes to the virulence of the pathogen by promoting its adherence to the mucosal surfaces [75,76]. It is also involved in the cell wall synthesis in MTb by catalysing the transesterification of mycolic acids [77].

The proteins of MTb and MAb, which do not form part of any orthologous cluster, are termed as "MTb-only" and "MAb-only" proteins, respectively, here, although some may have orthologs in other Mycobacterium species. These were assigned domains using HMMScan [78] against the profiles of all the families in the Pfam database [79]. Although these proteins from two mycobacterial species do not recognise each other in BLAST comparisons, some of these do possess domains in common. This non-recognition can be attributed to different domain organisation and evolving functions of related proteins. For example, the most frequent domain in the MAb-only set (11 times more as compared to its frequency in MTb-only) is TetR_N (tetracycline resistance repressor), which is a transcriptional regulator [80] (with HTH motif) and is known to regulate the degree of susceptibility of the microbe to hydrophobic

Table 1. Orthologous clusters in *Mycobacterium tuberculosis* and *Mycobacterium abscessus*. Three clusters with maximum number of genes in the orthologous cluster. The statistical attributes of the BLAST-searches (Expect value (E-value), query coverage (QC) and percentage identity (PID)) are listed in the table. E-values are used to report significance thresholds while performing similarity searches; they reflect the number of hits expected to occur by chance when searching for similar sequences in a given database of interest. The lower the E-value, the more significant the similarity. QC reflects the percent of the query sequence that aligns with the subject sequence and gives an estimate of the coverage of the alignment. PID is a measure of alignment quality and represents the percent of identical residues between query and subject sequence.

<i>Mycobacterium abscessus</i> proteins (UniProtKB IDs)		<i>Mycobacterium tuberculosis</i> proteins (UniProtKB IDs)					
		O53692 <i>ESAT-6-like protein EsxG</i>			Q6MX18 <i>ESAT-6-like protein EsxS</i>		
		E-value	QC (%)	PID (%)	E-value	QC (%)	PID (%)
Cluster I	B1MHU1 <i>PE family protein</i>	3.99e ⁻⁵⁰	99.0	75.0	5.83e ⁻⁴⁶	99.0	72.0
	B1ME84 <i>Probable PE family protein</i>	3.09e ⁻⁴⁷	99.0	73.9	4.47e ⁻⁴⁵	99.0	71.0
	B1MAN9 <i>Hypothetical PE family protein</i>	7.89e ⁻⁴⁹	99.0	76.1	1.18e ⁻⁴⁶	99.0	72.9
		P9WJV3 <i>Siderophore exporter MmpL4</i>					
		E-value	QC (%)	PID (%)			
Cluster II	B1MLC7 <i>Probable membrane protein, MmpL</i>	0.0	98.0	62.8			
	B1MLC6 <i>Probable membrane protein, MmpL</i>	0.0	98.0	63.3			
	B1MIG8 <i>Probable membrane protein, MmpL</i>	0.0	96.0	64.9			
	B1MIG7 <i>Probable membrane protein, MmpL</i>	0.0	98.0	63.5			
			P9WQN9 <i>DAG* acyltransferase/ mycolyltransferase Ag85C</i>				
					P9WQP1 <i>DAG* acyltransferase/ mycolyltransferase Ag85B</i>		
		E-value	QC (%)	PID (%)	E-value	QC (%)	PID (%)
Cluster III	B1MMV2 <i>Antigen 85-B</i>	8.15e ⁻¹³²	85.0	61.0	3.08e ⁻¹²³	97.0	56.8
	B1MEL3 <i>Antigen 85-A/B/C</i>	6.57e ⁻¹³⁰	85.0	60.7	1.16e ⁻¹²⁸	90.0	62.0
	B1MEL1 <i>Antigen 85-C</i>	4.66e ⁻¹³¹	85.0	63.1	4.66e ⁻¹³¹	85.0	65.1

*DAG: Diacylglycerol

antibiotics and detergents [81,82]. The MTb-only set was enriched in PE and PPE domains (~44 and 30 times more frequent as compared to MAb-only). The PE and PPE domains, believed to play immunological roles, contain a conserved N-terminal region and a more variable C-terminus [83–86]. Although 10% of MTb total genes are known to be part of these families of proteins [61], very few are structurally and functionally characterised.

This highlights that there is an urgent necessity for efforts to carry out the annotation of these poorly characterised genomes of infectious microorganisms, to tackle the infections and their drug resistance responses.

AVAILABLE INVENTORIES FOR MYCOBACTERIUM SPECIES

Since the complete genome sequence became available in 1998 [61], inventories for MTb genes have been developed [87,88]. The most recent version (Release 27) of Tuberculist [89] records the annotation of the whole genome of the H37Rv strain, and is one of the key resources available for the MTb genome. The tuberculosis

database (TBDB) documents the annotation and gene expression studies for multiple species and strains (including non-mycobacterial species), which enables one to perform comparative genomics [90]. Another useful and broad platform for analysing bacterial genomes and proteins is PATRIC (Pathosystems Resource Integration Center) [91], which allows the assembly and annotation of the bacterial genome(s) of interest, by providing tools for comparative genomics studies and RNA-seq analysis [91–93]. BioCyc has developed a specialised pathway tools website (<http://mycobacterium.biocyc.org/>), which contains 28 MTb genomes and provides a platform for genomic and metabolomic analyses and visualization.

The serious threat of drug resistance has led to efforts to build drug-resistance and drug-target databases, which aim to serve diagnostic roles and aid in the discovery of novel drugs and drug targets. The TB Drug Resistance Mutation Database (TBDReaMDB), developed in 2009, compiles a list of mutations against first and second line drugs used for treatment [94]. Another, recently developed, similar database, MUBII-TB-DB, documents the known drug-resistance-causing mutations at seven loci and enables the user

to search for these known mutations in the new sequence(s) of interest [95]. The TDR Targets database is a platform with a comprehensive collection of information about the gene products of tropical disease causing pathogens (including MTb) and allows the user to prioritize a target based on criteria such as structure, chemical interactions, and essentiality [96]. TuberQ is one of the databases that documents the essentiality and structural druggability of MTb proteins with known structures or high quality structural models [97].

STRUCTURE AND FUNCTION ANNOTATION OF MYCOBACTERIUM PROTEOME

For drug discovery and identification of new targets and drugs that can be used for treating MDR-TB (Multi-Drug Resistant TB) and XDR-TB (Extensively Drug Resistant TB), it becomes crucial to know more about the structure and function annotation of the MTb genome and its gene products. With the availability of the whole-genome sequence of MTb and more sensitive sequence search profile-based methods such as PSI-BLAST [98] and HMMER3 [78], it has become feasible to detect homologs (both close and distant) on the basis of both sequence and structure.

TubercuList maintains the most recent annotation of the MTb genome and classifies its proteins into eleven functional categories [89], such as regulatory proteins, proteins involved in cell wall and cell processes, and lipid metabolism. 27% of the MTb proteins lack annotation and are classified as either unknown or conserved hypothetical proteins. Assigning sequence and structure domains to these proteins might provide hints of protein function [99]. The search for structural templates using fold prediction [100-104] and homology modelling [105] methods enables one to build the structural model of a protein with unknown structure [106-109]. Using these combined strategies of homology detection will help in enriching the annotation and our knowledge about the functions and structure of the MTb proteome, which will further provide clues for structure-aided drug design. Holton *et al.* have aptly pressed the need to also focus on the structure of target complexes as well as single targets as these protein-protein interfaces can be targeted for drug design [110].

CHOPIN [109] is our in-house database built using an automated homology-modelling pipeline (*Vivace*) and stores homology models for 2911 MTb proteins (H37Rv

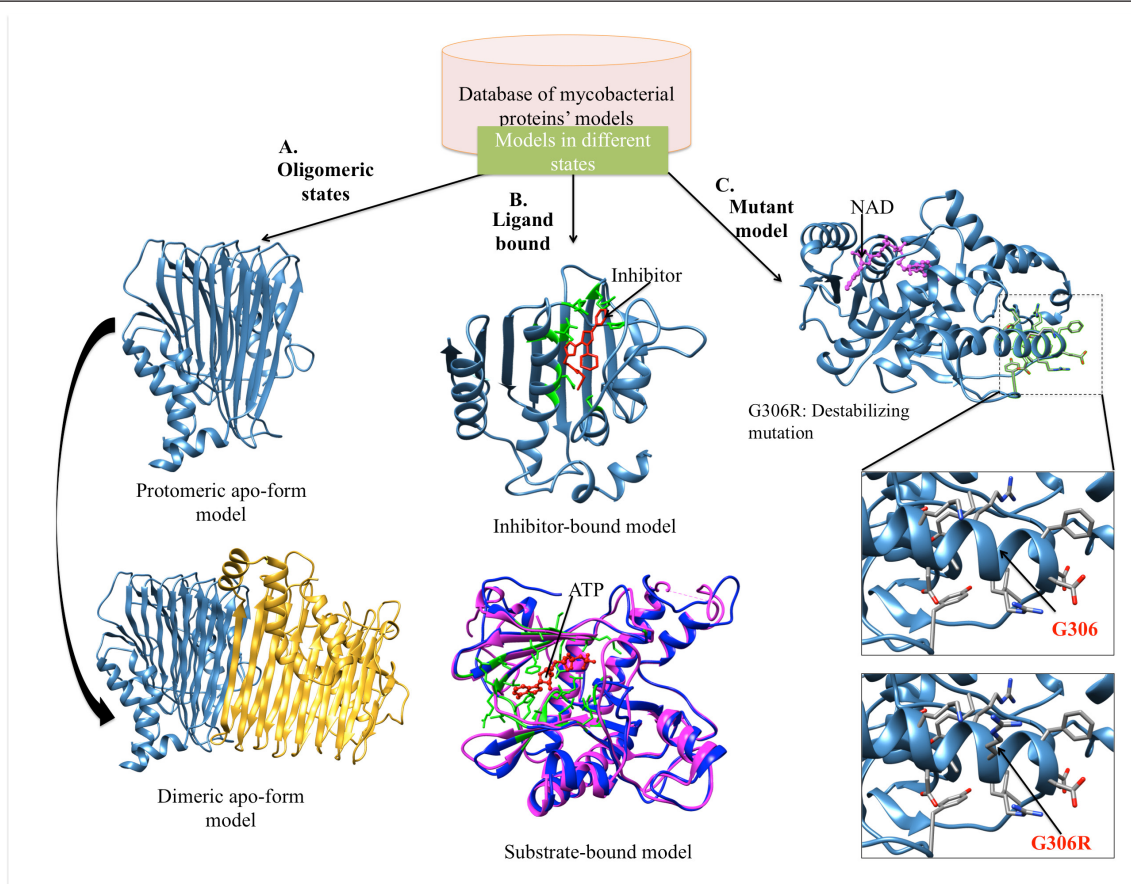


Figure 3. Structure and function annotation of *Mycobacterium* proteomes. The CHOPIN database of protein structure models for *Mycobacterium tuberculosis* proteome includes monomeric protein structures and also the ligand-bound structures (if a ligand-bound template can be used to build the model). **A.** An example shown from the database is the protomeric model of the *Mycobacterium tuberculosis* protein Rv1462 (conserved hypothetical protein of unknown function), built using the crystal structure of protein 1VH4 (stabilizer of iron transport). The oligomeric structures are now being implemented in the CHOPIN database for *Mycobacterium tuberculosis*. Here, the dimer built for the Rv1462 protein using the dimeric crystal structure of stabilizer of iron transport from *Escherichia coli* (PDB ID: 1VH4) is shown. **B.** Examples of ligand-bound structures in the database. Inhibitor-bound protein-structure model of the MTb DNA gyraseB (Rv0005), modeled using the inhibitor-bound *E. coli* DNA gyrase (PDB ID: 3G7E) as template. *Mycobacterium abscessus* protein ATCC19977_M00147132 (an enzyme involved in purine synthesis pathway, purC), modeled with ATP (shown in red) using the same pipeline as CHOPIN. In blue, is the homology model built using ATP-bound purC structure from *Saccharomyces cerevisiae* (PDB ID: 1OBD) and in pink is the recent apo-form crystal structure (PDB ID: 3R9R) of purC from *Mycobacterium abscessus*. **C.** Mutant model included in the CHOPIN database for the protein Rv1240 (malate dehydrogenase) having G306R mutation, which is known to have deleterious effect.

strain) (~73% coverage on the proteome) in different conformational states (apo and different ligand-bound). The modelling pipeline begins with identification of structural homolog(s) using a sequence-structure homology recognition approach, FUGUE [103]. This results in identification of a set of multiple templates that can be used for building a homology model using MODELER [105]. Figure 3 represents the different states of models (protomeric, oligomeric, inhibitor/substrate-bound and mutant models), which are included in the CHOPIN database. An example of annotation of an essential conserved protein with unknown function in MTb (Rv1462) is highlighted in figure 3A. Rv1462 possesses a Pfam domain (PF01458, residues 135-360), which represents sulphur assimilation (SUF) system of iron-sulphur cluster in eubacteria. The SUF system comprises SufABCDSE proteins forming SufSE and SufBCD assemblies. Rv1462 recognises a structural homolog (PDB ID: 1VH4 (SufD protein), stabiliser of iron transporter) with 25% identity upon Fugue searches and is modelled using our Fugue-Modeller pipeline in CHOPIN. The ligand coordinates from the template are also retained (if available) (Fig. 3B), which is advantageous, as modelling proteins in their functional states with other proteins and ligands allows one to predict their druggability.

CHOPIN also stores results of the predictions of effects of mutations, collated from TDRaMDB [94] and the Broad Institute [111], on protein stability using our in-house software – SDM and mCSM. Figure 3C shows a wild type and mutant model (G306R) of MTb protein Rv1240 (malate dehydrogenase enzyme). Site Directed Mutator, SDM [112,113] is based on the analysis of naturally occurring amino-acid substitutions and their use to predict thermodynamic quantities while mCSM -mutation Cut-off Scanning Matrix [114], is a machine learning approach, which relies on graph-based signatures, encoding distance patterns between atoms, used to represent the protein-residue environment and to train predictive models. Our recent analysis of mutations giving rise to resistance against front-line TB drugs has demonstrated a high frequency of mutations that affect allosteric sites and protein-protein interfaces in the proteins (Ascher D, Pires D and Blundell TL, unpublished data).

CHOPIN is part of the Structural Interactome Computational Resource (SinCRe) for MTb, which is an integrated database with domain architectures, functional and structural annotation, and protein-small-molecule interactions and their binding site information [115]. Recently, using profile-based search methods and protein fold prediction software, 95% of the MTb proteome (full or partial proteins) was assigned either functional or structural annotation [116]. Undoubtedly, these efforts are crucial but there is still a scope for further expansion of these resources. We are expanding CHOPIN to include protein structural models for additional MTb strains, information on protein-protein interactions and adopting more sensitive methods in the pipeline to increase the number of proteins that can be modelled (Malhotra S *et al.*, unpublished). We are also including the oligomeric models of MTb proteins based on the oligomeric status of homologs of known structure (Fig. 3A). We will assess the interfaces in the oligomeric models and score these assemblies (Malhotra S *et al.*, unpublished).

LESSONS FROM MYCOBACTERIUM GENOME

Given the challenge of treating MAb infections [56,64], progress in investigating and annotating the MTb genome has led to interest in examining the MAb genome. The MAb genome is very poorly structurally annotated, with only 30 structures available in the Protein Data Bank (PDB) [117]. We are now extending and optimising the tools and pipelines developed for the annotation of MTb genome to study its proteome, which should significantly improve our understanding of the Mab gene products and their functions. This will give insights into determining the folds and functions of the MAb proteins, as well as protein-protein and protein-ligand interactions (Malhotra S and Blundell TL, unpublished).

An example of this modelling pipeline adapted for modelling the MAb proteome is described in figure 3C. An enzyme from MAb is modelled using the ATP-bound purC structure from *Saccharomyces cerevisiae* (PDB ID: 1OBD). Recently, a crystal structure for MAb purC (apo) was also solved [118] (PDB ID: 3R9R, shown in pink in Fig. 3C). The modelled structure we built was structurally similar to the crystal structure and also had information about the ATP-binding residues. The ATP-binding pocket was also validated by superposing it over the crystal structure of ATP-bound crystal structure of the MAb purC enzyme (Thomas SE and Blundell TL, unpublished results).

TARGET IDENTIFICATION AND VALIDATION

Because genetic manipulation experiments of potential target genes are often challenging and time consuming, the preliminary choice of potential targets is usually based on previous studies establishing functional significance and essentiality of the target protein for bacterial growth and survival. If the protein has not been structurally or functionally characterized, building a structural model using homology modelling [105] can be useful. Mining the databases such as Tuberculist [89] and CHOPIN [109] can also provide important information regarding orthologs of the target in other species, its function and druggability. Target validation is often done by analysis of survival kinetics using conditional knockout (KO) strains where appropriate.

FRAGMENT-BASED DRUG DISCOVERY APPROACH - EXAMPLES AND IMPLICATIONS

In a recently published study from our group, we have described the application of a fragment-merging approach to develop small molecule inhibitors of *Mycobacterium tuberculosis* target protein EthR, in an effort to boost the effect of the existing second-line drug, Ethionamide [119]. This drug targets the enzyme 2-trans-enoyl reductase (InhA, part of type II fatty acid synthase) and is activated by a flavin-dependent monooxygenase, EthA, which is regulated by EthR. The small molecules that bind in the 20 Å hydrophobic cavity in EthR can be potentially used as booster compounds, as they possess an allosteric effect on its DNA-binding ability, which further weakens its transcriptional repressor activity. By merging the two fragments that bind in distinct regions in the hydrophobic cavity, the full cavity can be cov-

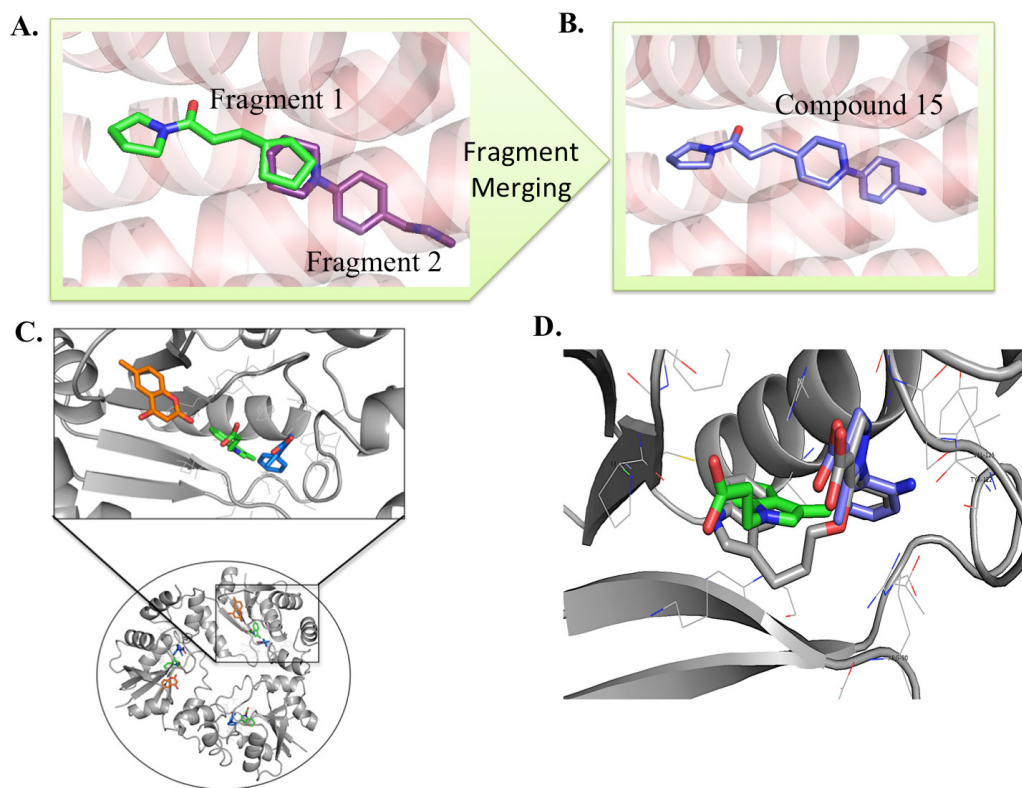


Figure 4. Fragment-merging approach to develop small-molecule inhibitors of *Mycobacterium tuberculosis* EthR. **A.** The region corresponding to the hydrophobic binding cavity is shown with the binding pocket of EthR where the starting fragment 1 (in green) and fragment 2 (in purple) were found to occupy distinct sites, thus filling the entire cavity. From [119], modified. **B.** Merging of the two fragments results in a significantly more potent lead (shown in blue) with $IC_{50} = 3 \mu M$. From [119], modified. **C.** A fragment-linking approach applied to phosphopantetheine adenylyltransferase (PPAT) protein from *Mycobacterium abscessus*. Superposition of three fragment-bound structures of PPAT protein (in grey) where fragments are shown in blue, green and orange, occupying three different regions of the protein's active site. **D.** PPAT with a molecule designed by linking strategies using the fragments shown in blue and green series occupying the adenylyl binding region of the active site.

ered (Fig. 4A) and the merged molecule binds EthR with increased potency ($IC_{50} = 3 \mu M$, Fig. 4B).

We are also exploring drug targets for MAb infection using the FBDD approach and have successfully applied it to phosphopantetheine adenylyltransferase (PPAT) protein, which catalyses the penultimate step in the biosynthesis of coenzyme-A in prokaryotes (Thomas SE *et al.*, unpublished results). PPAT protein consists of a large active site region that corresponds to sites for binding to 4-phosphopantetheine (substrate), the ribose phosphate moiety and adenylyl ring of 3'-dephospho-coenzyme-A (product), and coenzyme-A (feedback inhibitor). Fragments binding these three regions of the enzyme's active site were identified, as shown in figure 4C, D. Linking strategies on two such fragment series, each of which have affinities in mM ranges, have helped in the design of a compound with μM affinity.

CONCLUSIONS

The use of protein structure, defined by experimental and computational approaches for HIV protease in the 1980s and 1990s, for target identification and development of new leads, was an archetypal example of the use of structural biology to underpin drug discovery. The lessons learnt are proving helpful in developing structure-guided approaches to the design of antimicrobials against infectious diseases. Combined with fragment-based drug discovery, the ap-

proach allows efficient exploration of the chemical space against targets like those of mycobacteria, where features, such as the bacterial cell wall, demand the investigation of molecules that are likely different - in size, lipophilicity and other properties - than those optimised for human protein targets. The structure-guided fragment-based approach avoids the requirement for large screening libraries and provides a route to the discovery of new therapeutics that gives hope for less costly but effective medicines for rare diseases and those that affect the large populations of less affluent parts of the world.

REFERENCES

1. Insulin Crystals: Chemical & Biological Studies on Insulin Crystals & Insulin Zinc Suspensions by Schlichtkrull, Jorgen: Ejnar Munksgaard, Copenhagen - Expatriate Bookshop of Denmark
2. Adams MJ, Blundell TL, Dodson EJ, Dodson GG, Vijayan M, Baker EN, Harding MM, Hodgkin DC, Rimmer B, Sheat S (1969) Structure of Rhombohedral 2 Zinc Insulin Crystals. *Nature* 224: 491-495
3. Blundell TL, Cutfield JF, Cutfield SM, Dodson EJ, Dodson GG, Hodgkin DC, Mercola DA, Vijayan M (1971) Atomic positions in rhombohedral 2-zinc insulin crystals. *Nature* 231: 506-511
4. Browne WJ, North ACT, Phillips DC, Brew K, Vanaman TC, Hill RL (1969) A possible three-dimensional structure of bovine α -lactalbumin based on that of hen's egg-white lysozyme. *J Mol Biol* 42: 65-86
5. Bedarkar S, Turnell WG, Blundell TL, Schwabe C (1977) Relaxin has conformational homology with insulin. *Nature* 270: 449-451
6. Blundell TL, Bedarkar S, Rinderknecht E, Humbel RE (1978) Insulin-like growth factor: a model for tertiary structure accounting for

- immunoreactivity and receptor binding. *Proc Natl Acad Sci U S A* 75: 180–184
7. Blundell T (1979) Conformation and molecular biology of polypeptide hormones I. Insulin, insulin-like growth factor and relaxin. *Trends Biochem Sci* 4: 51–54
 8. Blundell TL, Humbel RE (1980) Hormone families: pancreatic hormones and homologous growth factors. *Nature* 287: 781–787
 9. Jones TA (1978) A graphics model building and refinement system for macromolecules. *J Appl Crystallogr* 11: 268–272
 10. Blundell T, Sibanda BL, Pearl L (1983) Three-dimensional structure, specificity and catalytic mechanism of renin. *Nature* 304: 273–275
 11. Pearl L, Blundell T (1984) The active site of aspartic proteinases. *FEBS Lett* 174: 96–101
 12. Rahuel J, Priestle JP, Grütter MG (1991) The crystal structures of recombinant glycosylated human renin alone and in complex with a transition state analog inhibitor. *J Struct Biol* 107: 227–236
 13. Bott R, Subramanian E, Davies DR (1982) Three-dimensional structure of the complex of the *Rhizopus chinensis* carboxyl proteinase and pepstatin at 2.5-Å resolution. *Biochemistry (Mosc)* 21: 6956–6962
 14. Veerapandian B, Cooper JB, Sali A, Blundell TL (1990) X-ray analyses of aspartic proteinases. III Three-dimensional structure of endothiapepsin complexed with a transition-state isostere inhibitor of renin at 1.6 Å resolution. *J Mol Biol* 216: 1017–1029
 15. James MN, Sielecki AR (1986) Molecular structure of an aspartic proteinase zymogen, porcine pepsinogen, at 1.8 Å resolution. *Nature* 319: 33–38
 16. Argos P (1976) Prediction of the secondary structure of mouse nerve growth factor and its comparison with insulin. *Biochem Biophys Res Commun* 70: 805–811
 17. Wlodawer A, Hodgson KO, Shooter EM (1975) Crystallization of nerve growth factor from mouse submaxillary glands. *Proc Natl Acad Sci U S A* 72: 777–779
 18. McDonald NQ, Lapatto R, Murray-Rust J, Gunning J, Wlodawer A, Blundell TL (1991) New protein fold revealed by a 2.3-Å resolution crystal structure of nerve growth factor. *Nature* 354: 411–414
 19. Tang J, James MN, Hsu IN, Jenkins JA, Blundell TL (1978) Structural evidence for gene duplication in the evolution of the acid proteases. *Nature* 271: 618–621
 20. Toh H, Ono M, Saigo K, Miyata T (1985) Retroviral protease-like sequence in the yeast transposon Ty 1. *Nature* 315: 691–691
 21. Pearl LH, Taylor WR (1987) A structural model for the retroviral proteases. *Nature* 329: 351–354
 22. Miller M, Jaskólski M, Rao JK, Leis J, Wlodawer A (1989) Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature* 337: 576–579
 23. Navia MA, Fitzgerald PMD, McKeever BM, Leu C-T, Heimbach JC, Herber WK, Sigal IS, Darke PL, Springer JP (1989) Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature* 337: 615–620
 24. Blundell T, Pearl L (1989) A second front against AIDS. *Nature* 337: 596–597
 25. Lapatto R, Blundell T, Hemmings A, Overington J, Wilderspin A, Wood S, Merson JR, Whittle PJ, Danley DE, Geoghegan KF (1989) X-ray analysis of HIV-1 proteinase at 2.7 Å resolution confirms structural homology among retroviral enzymes. *Nature* 342: 299–302
 26. Li M, Gustchina A, Cruz R, Simões M, Curto P, Martinez J, Faro C, Simões I, Wlodawer A (2015) Structure of RC1339/APRc from *Rickettsia conorii*, a retropepsin-like aspartic protease. *Acta Crystallogr D Biol Crystallogr* 71: 2109–2118
 27. Gustchina A, Li M, Phylip LH, Lees WE, Kay J, Wlodawer A (2002) An unusual orientation for Tyr75 in the active site of the aspartic proteinase from *Saccharomyces cerevisiae*. *Biochem Biophys Res Commun* 295: 1020–1026
 28. Li M, Phylip LH, Lees WE, Winther JR, Dunn BM, Wlodawer A, Kay J, Gustchina A (2000) The aspartic proteinase from *Saccharomyces cerevisiae* folds its own inhibitor into a helix. *Nat Struct Biol* 7: 113–117
 29. Bryksa BC, Bhaumik P, Magracheva E, De Moura DC, Kurylowicz M, Zdanov A, Dutcher JR, Wlodawer A, Yada RY (2011) Structure and mechanism of the saposin-like domain of a plant aspartic protease. *J Biol Chem* 286: 28265–28275
 30. Kervinen J, Tobin GJ, Costa J, Waugh DS, Wlodawer A, Zdanov A (1999) Crystal structure of plant aspartic proteinase prophylpsin: inactivation and vacuolar targeting. *EMBO J* 18: 3947–3955
 31. Li M, Gustchina A, Matúz K, Tözsér J, Namwong S, Goldfarb NE, Dunn BM, Wlodawer A (2011) Structural and biochemical characterization of the inhibitor complexes of xenotropic murine leukemia virus-related virus protease. *FEBS J* 278: 4413–4424
 32. Satoh T, Li M, Nguyen J-T, Kiso Y, Gustchina A, Wlodawer A (2010) Crystal structures of inhibitor complexes of human T-cell leukemia virus (HTLV-1) protease. *J Mol Biol* 401: 626–641
 33. Li M, Dimairo F, Zhou D, Gustchina A, Lubkowski J, Dauter Z, Baker D, Wlodawer A (2011) Crystal structure of XMRV protease differs from the structures of other retropepsins. *Nat Struct Mol Biol* 18: 227–229
 34. Powell DJ, Bur D, Wlodawer A, Gustchina A, Dunn BM, Kay J (1998) The aspartic proteinase from equine infectious anaemia virus. *Adv Exp Med Biol* 436: 41–45
 35. Abagyan R, Totrov M (2001) High-throughput docking for lead generation. *Curr Opin Chem Biol* 5: 375–382
 36. Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28: 849–857
 37. Böhm HJ (1992) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* 6: 593–606
 38. Gillet VJ, Myatt G, Zsoldos Z, Johnson AP (1995) SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect Drug Discov Des* 3: 34–50
 39. Carlson HA, Masukawa KM, Rubins K, Bushman FD, Jorgensen WL, Lins RD, Briggs JM, McCammon JA (2000) Developing a dynamic pharmacophore model for HIV-1 integrase. *J Med Chem* 43: 2100–2114
 40. Blundell TL, Jhoti H, Abell C (2002) High-throughput crystallography for lead discovery in drug design. *Nat Rev Drug Discov* 1: 45–54
 41. Miranker A, Karplus M (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* 11: 29–34
 42. Verdonk ML, Cole JC, Taylor R (1999) SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J Mol Biol* 289: 1093–1108
 43. Verdonk ML, Cole JC, Watson P, Gillet V, Willett P (2001) Superstar: improved knowledge-based interaction fields for protein binding sites I. *J Mol Biol* 307: 841–859
 44. Radoux CJ, Olsson TSG, Pitt WR, Groom CR, Blundell TL (2016) Identifying interactions that determine fragment binding at protein hotspots. *J Med Chem* 59: 4314–4325
 45. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein–ligand docking using GOLD. *Proteins Struct Funct Bioinforma* 52: 609–623
 46. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* 30: 2785–2791
 47. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *J Med Chem* 49: 6177–6196
 48. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* 47: 1750–1759
 49. Joseph-McCarthy D, Campbell AJ, Kern G, Moustakas D (2014) Fragment-based lead discovery and design. *J Chem Inf Model* 54: 693–704
 50. Murray CW, Verdonk ML, Rees DC (2012) Experiences in fragment-based drug discovery. *Trends Pharmacol Sci* 33: 224–232

51. Scott DE, Coyne AG, Hudson SA, Abell C (2012) Fragment-based approaches in drug discovery and chemical biology. *Biochemistry (Mosc)* 51: 4990–5003
52. Hubbard RE, Murray JB (2011) Experiences in fragment-based lead discovery. *Methods Enzymol* 493: 509–531
53. Winter A, Sigurdardottir AG, DiCara D, Valenti G, Blundell TL, Gherardi E (2016) Developing antagonists for the Met-HGF/SF protein-protein interaction using a fragment-based approach. *Mol Cancer Ther* 15: 3–14
54. Mycobacterium (<http://www.bacterio.net/mycobacterium.html>)
55. WHO, WHO | Global tuberculosis report 2015
56. Nessar R, Cambau E, Reyrat JM, Murray A, Gicquel B (2012) Mycobacterium abscessus: a new antibiotic nightmare. *J Antimicrob Chemother* 67:810-8
57. Griffith DE, Aksamit T, Brown-Elliott BA, Catanzaro A, Daley C, Gordin F, Holland SM, Horsburgh R, Huitt G, Iademarco MF, et al. (2007) An Official ATS/IDSA statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. *Am J Respir Crit Care Med* 175: 367–416
58. Lee M-R, Sheng W-H, Hung C-C, Yu C-J, Lee L-N, Hsueh P-R (2015) *Mycobacterium abscessus* complex infections in humans. *Emerg Infect Dis* 21: 1638–1646
59. Bioproject Grid (<https://www.ncbi.nlm.nih.gov/bioproject/browse/>)
60. Sreevatsan S, Pan XI, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci* 94: 9869–9874
61. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537–544
62. Cole ST (1999) Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett* 452: 7–10
63. Moore M, Frerichs JB (1953) An unusual acid-fast infection of the knee with subcutaneous, abscess-like lesions of the gluteal region; report of a case with a study of the organism, *Mycobacterium abscessus*, n. sp. *J Invest Dermatol* 20: 133–169
64. Ripoll F, Pasek S, Schenowitz C, Dossat C, Barbe V, Rottman M, Macheras E, Heym B, Herrmann J-L, Daffé M, et al. (2009) Non mycobacterial virulence genes in the genome of the emerging pathogen *Mycobacterium abscessus*. *PLOS ONE* 4: e5660
65. The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43: D204–D212
66. Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ (2011) Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* 12: 124
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410
68. Tufariello JM, Chapman JR, Kerantzas CA, Wong K-W, Vilchèze C, Jones CM, Cole LE, Tinaztepe E, Thompson V, Fenyö D, et al. (2016) Separable roles for *Mycobacterium tuberculosis* ESX-3 effectors in iron acquisition and virulence. *Proc Natl Acad Sci U S A* 113: E348–357
69. Arbing MA, Kaufmann M, Phan T, Chan S, Cascio D, Eisenberg D (2010) The crystal structure of the *Mycobacterium tuberculosis* Rv3019c-Rv3020c ESX complex reveals a domain-swapped heterotetramer. *Protein Sci Publ Protein Soc* 19: 1692–1703
70. Fischer K, Chatterjee D, Torrelles J, Brennan PJ, Kaufmann SH, Schaike UE (2001) Mycobacterial lysocardiolipin is exported from phagosomes upon cleavage of cardiolipin by a macrophage-derived lysosomal phospholipase A2. *J Immunol Baltim Md* 167: 2187–2192
71. Domenech P, Reed MB, Barry CE (2005) Contribution of the *Mycobacterium tuberculosis* MmpL protein family to virulence and drug resistance. *Infect Immun* 73: 3492–3501
72. Bailo R, Bhatt A, Ainsa JA (2015) Lipid transport in *Mycobacterium tuberculosis* and its implications in virulence and drug development. *Biochem Pharmacol* 96: 159–167
73. Pasca MR, Gugliera P, Rossi ED, Zara F, Riccardi G (2005) mmpL7 Gene of *Mycobacterium tuberculosis* is responsible for isoniazid efflux in *Mycobacterium smegmatis*. *Antimicrob Agents Chemother* 49: 4775–4777
74. Székely R, Cole ST (2016) Mechanistic insight into mycobacterial MmpL protein function. *Mol Microbiol* 99: 831–834
75. Wiker HG, Harboe M (1992) The antigen 85 complex: a major secretion product of *Mycobacterium tuberculosis*. *Microbiol Rev* 56: 648–661
76. Abou-Zeid C, Garbe T, Lathigra R, Wiker HG, Harboe M, Rook GA, Young DB (1991) Genetic and immunological analysis of *Mycobacterium tuberculosis* fibronectin-binding proteins. *Infect Immun* 59: 2712–2718
77. Belisle JT, Vissa VD, Sievert T, Takayama K, Brennan PJ, Besra GS (1997) Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. *Science* 276: 1420–1422
78. Eddy SR (2011) Accelerated profile HMM searches. *PLOS Comput Biol* 7: e1002195
79. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44: D279–D285
80. Hillen W, Berens C (1994) Mechanisms underlying expression of Tn10 encoded tetracycline resistance. *Annu Rev Microbiol* 48: 345–369
81. Pan W, Spratt BG (1994) Regulation of the permeability of the gonococcal cell envelope by the mtr system. *Mol Microbiol* 11: 769–775
82. Ramos JL, Martínez-Bueno M, Molina-Henares AJ, Terán W, Watanabe K, Zhang X, Gallegos MT, Brennan R, Tobes R (2005) The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* 69: 326–356.
83. Brennan MJ, Delogu G (2002) The PE multigene family: a ‘molecular mantra’ for mycobacteria. *Trends Microbiol* 10: 246–249
84. Sultana R, Tanneeru K, Kumar ABR, Guruprasad L (2016) Prediction of certain well-characterized domains of known functions within the PE and PPE proteins of Mycobacteria. *PLOS ONE* 11: e0146786
85. Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE (2012) The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: perhaps more? *Biochimie* 94: 110–116
86. Fishbein S, van Wyk N, Warren RM, Sampson SL (2015) Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Mol Microbiol* 96: 901–916
87. Ekins S, Freundlich JS, Choi I, Sarker M, Talcott C (2011) Computational databases, pathway and cheminformatics tools for tuberculosis drug discovery. *Trends Microbiol* 19: 65–74
88. Lew JM, Mao C, Shukla M, Warren A, Will R, Kuznetsov D, Xenarios I, Robertson BD, Gordon SV, Schnappinger D, et al. (2013) Database resources for the tuberculosis community. *Tuberc Edinb Scotl* 93: 12–17
89. Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList – 10 years after. *Tuberc Edinb Scotl* 91: 1–7
90. Galagan JE, Sisk P, Stolte C, Weiner B, Koehrsen M, Wymore F, Reddy TBK, Zucker JD, Engels R, Gellesch M, et al. (2010) TB database 2010: overview and update. *Tuberc Edinb Scotl* 90: 225–235
91. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, et al. (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 42: D581–591
92. Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, et al. (2011) PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun* 79: 4286–4298
93. Wattam AR, Gabbard JL, Shukla M, Sobral BW (2014) Comparative genomic analysis at the PATRIC, a bioinformatic resource center. *Methods Mol Biol Clifton NJ* 1197: 287–308
94. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB (2009) Tuberculosis drug resistance mutation database. *PLOS Med* 6: e1000002

95. Flandrois J-P, Lina G, Dumitrescu O (2014) MUBII-TB-DB: a database of mutations associated with antibiotic resistance in *Mycobacterium tuberculosis*. BMC Bioinformatics 15: 107
96. Magariños MP, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, Shanmugam D, Van Voorhis WC, Agüero F (2012) TDR Targets: a chemogenomics resource for neglected diseases. Nucleic Acids Res 40: D1118-D1127
97. Radusky L, Defelipe LA, Lanzarotti E, Luque J, Barril X, Marti MA, Turjanski AG (2014) TuberQ: a *Mycobacterium tuberculosis* protein druggability database. Database 2014: bau035
98. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402
99. Namboori S, Mhatre N, Sujatha S, Srinivasan N, Pandit SB (2004) Enhanced functional and structural domain assignments using remote similarity detection procedures for proteins encoded in the genome of *Mycobacterium tuberculosis* H37Rv. J Biosci 29: 245-259
100. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10: 845-858
101. Hildebrand A, Remmert M, Biegert A, Söding J (2009) Fast and accurate automatic structure prediction with HHpred. Proteins 77 Suppl 9: 128-132
102. Yachdav G, Koppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Hönigschmid P, Schafferhans A, Roos M, Bernhofer M, et al. (2014) PredictProtein – an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 42: W337-343
103. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310: 243-257
104. Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287: 797-815
105. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779-815
106. Mao C, Shukla M, Larrouy-Maumus G, Dix FL, Kelley LA, Sternberg MJ, Sobral BW, de Carvalho LPS (2013) Functional assignment of *Mycobacterium tuberculosis* proteome revealed by genome-scale fold-recognition. Tuberc Edinb Scotl 93: 40-46
107. Chandra N, Sandhya S, Anand P (2014) Structural annotation of the *Mycobacterium tuberculosis* proteome. Microbiol Spectr 2: 261-280
108. Anand P, Sankaran S, Mukherjee S, Yeturu K, Laskowski R, Bhardwaj A, Bhagavat R, OSDD Consortium, Brahmachari SK, Chandra N (2011) Structural annotation of *Mycobacterium tuberculosis* proteome. PloS One 6: e27044
109. Ochoa-Montaña B, Mohan N, Blundell TL (2015) CHOPIN: a web resource for the structural and functional proteome of *Mycobacterium tuberculosis*. Database 2015: bav026
110. Holton SJ, Weiss MS, Tucker PA, Wilmanns M (2007) Structure-based approaches to drug discovery against tuberculosis. Curr Protein Pept Sci 8: 365-375
111. *Mycobacterium tuberculosis* Comparative – Olive (<https://olive.broadinstitute.org/projects/Mycobacterium%2520tuberculosis%-2520Comparative>)
112. Worth CL, Preissner R, Blundell TL (2011) SDM – a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 39: W215-222
113. Topham CM, Srinivasan N, Blundell TL (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng 10: 7-21
114. Pires DEV, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30: 335-342
115. Metri R, Hariharaputran S, Ramakrishnan G, Anand P, Raghavender US, Ochoa-Montaña B, Higuero AP, Sowdhamini R, Chandra NR, Blundell TL, et al. (2015) SInCRE-structural interactome computational resource for *Mycobacterium tuberculosis*. Database J Biol Databases Curation 2015: bav060
116. Ramakrishnan G, Ochoa-Montaña B, Raghavender US, Mudgal R, Joshi AG, Chandra NR, Sowdhamini R, Blundell TL, Srinivasan N (2015) Enriching the annotation of *Mycobacterium tuberculosis* H37Rv proteome using remote homology detection approaches: insights into structure and function. Tuberc Edinb Scotl 95: 14-25
117. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242
118. Baugh L, Phan I, Begley DW, Clifton MC, Armour B, Dranow DM, Taylor BM, Muruthi MM, Abendroth J, Fairman JW, et al. (2015) Increasing the structural coverage of tuberculosis drug targets. Tuberc Edinb Scotl 95: 142-148
119. Nikiforov PO, Surade S, Blaszczyk M, Delorme V, Brodin P, Baulard AR, Blundell TL, Abell C (2016) A fragment merging approach towards the development of small molecule inhibitors of *Mycobacterium tuberculosis* EthR for use as ethionamide boosters. Org Biomol Chem 14: 2318-2326
120. Freed EO (2015) HIV-1 assembly, release and maturation. Nat Rev Microbiol 13: 484-496

Wspomagane strukturą ukierunkowane projektowanie leku – eksploracja informacji genetycznej od wirusa HIV po mykobakterie

Sony Malhotra, Sherine E. Thomas, Bernardo Ochoa Montano, Tom L. Blundell✉

Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, UK

✉e-mail: tlb20@hermes.cam.ac.uk or tom@cryst.bioc.cam.ac.uk

Słowa kluczowe: wspomagane strukturą, oparte na fragmencie, projektowanie/o3939997dkrywanie leku, proteaza HIV, inhibitory, *Mycobacterium tuberculosis*, *Mycobacterium abscessus*

STRESZCZENIE

Zastosowanie krystalografii białek podczas wspomaganej strukturą opracowywania leków umożliwia identyfikację potencjalnych miejsc wiązania inhibitora oraz optymalizację warunków oddziaływań pomiędzy związkami wiodącymi a białkiem docelowym. Przykładem wcześniejszego zastosowania tego typu podejścia było użycie struktury proteazy HIV podczas projektowania leków antywirusowych przeciw AIDS. W ostatnim czasie wspomagane strukturą opracowywanie leku pozwoliło zwiększyć wydajność wiązania liganda, a dzięki oparciu projektowania na fragmencie cząsteczki zmniejszono złożoność całego procesu i wielkość biblioteki selekcyjnej. W niniejszym artykule omówiono zastosowanie metody wspomaganej strukturą identyfikacji związku docelowego oraz optymalizacji związku wiodącego podczas opartej na fragmencie opracowywania nowych związków antybakteryjnych do stosowania w leczeniu zakażeń mykobakteriami.