Chapter

Computational Deorphaning of *Mycobacterium tuberculosis* Targets

Lorraine Yamurai Bishi, Sundeep Chaitanya Vedithi, Tom L. Blundell and Grace Chitima Mugumbate

Abstract

Tuberculosis (TB) continues to be a major health hazard worldwide due to the resurgence of drug discovery strains of *Mycobacterium tuberculosis* (*Mtb*) and co-infection. For decades drug discovery has concentrated on identifying ligands for ~10 *Mtb* targets, hence most of the identified essential proteins are not utilised in TB chemotherapy. Here computational techniques were used to identify ligands for the orphan *Mtb* proteins. These range from ligand-based and structure-based virtual screening modelling the proteome of the bacterium. Identification of ligands for most of the *Mtb* proteins will provide novel TB drugs and targets and hence address drug resistance, toxicity and the duration of TB treatment.

Keywords: *Mycobacterium tuberculosis*, target deorphaning, target deconvolution, proteome modelling, virtual screening

1. Introduction

Tuberculosis (TB) continues to be a major public health concern with over 2 billion people currently infected, 8.6 million new cases per year, and more than 1.3 million deaths annually [1]. The current drug-regimen combination for drug sensitive TB consists of isoniazid, rifampicin, ethambutol and pyrazinamide, administered over 6 months [2]. If this treatment fails, second-line drugs are used, such as para-aminosalicylate (PAS) and fluoroquinolones, which are usually either less effective or more toxic with serious side effects. Although this regimen has a high success rate, it is marred by compliance issues, which have resulted in the rise of multidrug resistant (MDR), extensively drug resistant (XDR) and totally drug resistant (TDR) strains of the causative agent, *Mycobacterium tuberculosis* (*Mtb*) [3, 4], in both immunocompetent and immunocompromised patients worldwide [5]. However, it took about 40 years for a new TB drug to be discovered and most of the current TB drugs target a total of only ~10 proteins, even though the complete genome of *Mtb* was published nearly 20 years ago [6]. Consequently, most of the essential proteins are orphans since their ligands are still to be identified. In our context, target deorphaning or deconvolution encompasses identification of ligands for *Mtb* proteins not currently exploited in TB chemotherapy and those of old TB targets. Targeting further essential proteins should allow the fight against drug resistance to be enhanced, and possibly lead to a reduction in the duration of TB treatment.

The conventional target deorphaning process involves experimental work, which characteristically includes genetic, proteomics and transcriptional profiling and then identification of the ligands for the proteins using many more chemical-proteomic approaches [7]. This approach is usually long, expensive and time consuming. However, developments in bioinformatics and chemoinformatics, together with advances in computer tools and resources, have fortunately revolutionised target deorphaning. Bioinformatics describes the target space in *Mtb* from the genome to the proteome, whilst chemoinformatics provides information about the available chemical space and tools for navigation of the space. Together these developments have led to a mushrooming of computer-based target deorphaning methods ranging from modelling proteomes, virtual screening, machine and deep learning, and chemogenomics [8–10]. When used effectively in conjunction with experimental work, computational methods can facilitate identification of new TB targets and drugs [11–13].

Therefore, in this chapter we present an overview of the genome of *Mtb*, giving a detailed account on how the computational techniques have been used to de-orphan *Mtb* targets including case studies, the current and proposed future impacts of these techniques on the number of de-orphaned *Mtb* targets and their impacts in boosting the biomedical efficacy of TB drugs. The collated data will provide researchers in academia and industry with knowledge of target-ligand pairs and interactions, information crucial for the design of novel drugs with known targets that are less prone to resistance, with minimal side effects and interactions with e.g. anti-HIV drugs.

2. Method

An extensive literature search was performed to give an overview of the genome of the *Mtb* and status of the currently used tuberculosis drugs and their targets. An analysis of the essential proteins in *Mtb* and the number of proteins targeted by the current TB drugs was performed. To boost this data *Mtb* target-ligand data was extracted from the ChEMBL database version 24 (https://www.ebi.ac.uk/chembl/beta/g/#browse/targets), which was used to determine the number of the proposed new targets. An overview of computational deorphaning of *Mtb* targets is provided, using data extracted from literature and a description of the efforts made from our laboratory. To sum this up, a detailed account of modelling the proteome for Mycobacteria, and identification of the hotspots and druggability of the proteins is given.

3. Genome sequence of *Mycobacterium tuberculosis*

Cole and co-workers [14] in 1998 reported the complete sequence of *Mtb,* which comprises of 4,411,529 base pairs. The genome has an evenly distributed guanine-cysteine content of 65.6% and represents the second-largest bacterial genome sequence currently available. Additionally, the genome is rich in repetitive DNA, particularly insertion sequences, and in new multi-gene families and duplicated housekeeping genes, providing evidence for horizontally-transferred pathogenicity islands of a particular base composition [14].

The genome of *Mtb* has some exceptional features, for example there are over 200 genes that encode enzymes for the metabolism of fatty acids, comprising 6% of the total (**Table 1**). Among these, about 100 are predicted to function in the oxidation of fatty acids. This large number of *Mtb* enzymes that putatively have fatty acids as substrates may be linked to the ability of this pathogen to grow in the tissues of the infected host, where fatty acids maybe the major carbon source. Another

Function	No. of genes	% of total genes	% of total coding capacity
Lipid metabolism	225	5.7	9.3
Information pathways	207	5.2	6.1
Cell wall and cell processes	517	13.0	13.5
Stable RNAs	50	1.3	0.2
IS elements and bacteriophages	137	3.4	2.5
PE and PPE Proteins	167	4.2	7.1
Intermediary metabolism and respiration	877	22.0	24.6
Regulatory proteins	188	4.7	4.0
Virulence, detoxification and adaptation	91	2.3	2.4
Conserved hypothetical function	911	22.9	18.4
Proteins of unknown function	607	15.3	9.9

Table 1.
General classification of Mtb *genes.* Adopted from *[15]*.

unusual feature of the *Mtb* genome is the presence of the unrelated Pro-Glu (PE) and Pro-Pro-Glu (PPE) families of proteins that have conserved N-terminal domains of 100 and 180 amino acids respectively. The antigenicity of these proteins has led to the assumption that at least some of these proteins may be involved in antigenic variation of *Mtb* during infection [15].

3.1 Current status of tuberculosis drugs and targets

3.1.1 Tuberculosis drugs

The success of TB chemotherapy derives from an "intensive" phase involving a cocktail of four first-line drugs, comprising, rifampicin (RIF), isoniazid (INH), pyrazinamide (PZA), and ethambutol (EMB). A threatening global issue of this epidemic is the emergence of drug-resistant bacteria, a trend that is on the rise, as such strains are easily spread with low fitness costs associated with transmission [16]. The World Health Organisation (WHO) reported that globally 3.5% of naive infections already expressed resistance to the two most efficacious frontline agents used to treat the disease, RIF and INH, thereby classifying the infection as multi-drug resistant tuberculosis (MDR-TB) [17]. Treatment of drug-resistant *Mtb* is difficult already, requiring 6–9 months of combination therapy of second-line drugs, such as PAS, fluoroquinolones e.g. levofloxacin, and aminoglycosides e.g. kanamycin, capreomycin, ethionamide and cycloserine. Complicating the issue is the fact that TB is endemic to the developing world; thus, access to adequate healthcare facilities and drugs can be limited for those patients. This leads to non-compliance by most patients, relapse of the disease and severe side-effects especially of second-line drugs [18]. Treatment for MDR-TB can extend upwards of 2 years and relies on more toxic, less efficacious second-line drugs, many of which are even more scarce than frontline drugs in affected areas [16].

In addition, comorbidity with HIV causes massive diagnostic and therapeutic challenges and results in adverse drug interactions [19]. This is because RIF is a potent inducer of drug-metabolising enzymes, including cytochrome P450 (CYP) 3A4. This induction dramatically reduces plasma levels of several highly active antiretroviral therapy drugs; thus, patients are often forced to complete

their TB treatment before beginning HIV treatment [20]. Patients who contract MDR-TB with HIV have a very poor prognosis due to the duration of treatment; these individuals frequently succumb within a few months. Therefore, there is an urgent need to develop continually new active agents to combat MDR-TB which has been compounded by the emergence of XDR-TB. Furthermore, cases of TDR-TB have been noted in China, India, Africa, and Eastern Europe. In TDR-TB, the *Mycobacterium* are resistant to all available therapeutics [19]. To address this, in 2012 the U.S. Food and Drug Agency (FDA) approved bedaquiline for MDR-TB [21] and later delamanid was approved as a compassionate care option for XDR-TB and TDR-TB infections, nonetheless the EMA approved both agents for MDR-TB [22]. The biggest challenge is that these drugs have reported human ether-a-go-go related gene (hERG) toxicity, as well as multiple absorption, distribution, metabolism and excretion (ADME) issues due to their high lipophilicity [21]. This leads to an urgent need for development of new agents that have successful therapeutic effects.

3.1.2 Mycobacterium tuberculosis drug targets

To date the number of essential *Mtb* proteins encoded by approximately 4000 genes is just over 500 (**Figure 1**), and this provides a rich source for novel targets for new and current TB drugs. However, Lamichhane et al. [23] reported that TB chemotherapy exploited only 10 of these proteins; **Table 2**, gives a summary of the targets, and their current and/or new drug ligands. The most popular target is enoyl[acyl-carrier protein] reductase, important for the biosynthesis of mycolic acid. Efforts to identify genes that code for new potential drugs are underway, as evidenced by 76 TB data points recorded in the ChEMBL database version 24 (https://www.ebi.ac.uk/chembl/beta/g/#browse/targets), consisting of small bioactive compounds, their targets and bioassay data. There are 73 single proteins, including the 10 proteins already targeted by both first-line and second-line drugs during TB chemotherapy. Thus, 63 new drug targets are being explored in a plethora of bioassays.

This is of paramount importance because *Mtb* secreted proteins play a vital role in host-pathogen interactions and facilitate nutrient acquisition, pilot the host immune response and interfere with therapeutic intervention. Therefore, the *Mtb* secretome consists of proteins essential for successful invasion and *in vivo* growth during host infection. The essential proteins are the most suitable drug targets for the development of diagnostic tools and new drugs, because of their key role in *in vivo* bacterial survival and growth. Identifying ligands for these proteins required for growth and survival in the infected host could lead to the discovery of potentially useful biomarkers to add on the above mentioned drug targets [27].

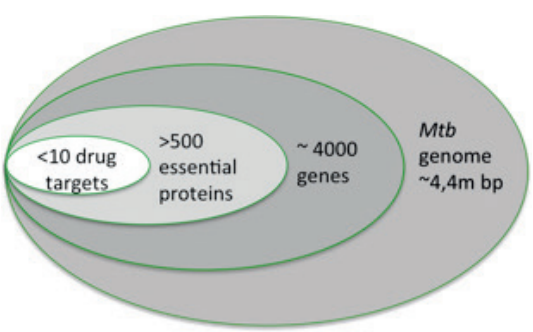

Figure 1.
Circular diagram of the genome of Mtb *genes, essential proteins and the number of proteins that are drug targets.*

Targets	Function	Conventional drugs	New ligands
Enoyl-(acyl-carrier-protein) reductase (InhA), Fatty acid synthase	Biosynthesis of mycolic acids, that is essential for growth and virulence	Isoniazid Ethambutol Pyrazinamide Delamanid	Tetrahydropyrans (PT070) Methylthiazoles Diazaborines Pyrrolidine-carboxamide Piperazine indoleformamides Aminoproline Arylamides Imidazopiperidines
DNA gyrase	An ATP-dependent enzyme that acts by creating a transient double-stranded DNA break	Fluoroquinolones	Clinafloxacin
Ubiquinol-cytochrome C-reductase (QCrB)	Electron carriers of the respiratory chain		Pyrrolo[3,4-c]pyridine-1,3(2H)diones Lansoprazole
Transmembrane transport protein large (MmpL3)	Responsible for heme uptake into the cell. Responsible for the transport of ions, drugs, fatty acids and bile salts		SQ109 Adamantyl ureas Phenylpyrroles Benzimidazoles Tetrahydropyrazolo [1,5-a]pyrimidine-3-carboxamide Spiropiperidines
Decaprenylphospo-β-D-ribofuranose-2-oxidase (DprE1)	Cell wall synthesis		Benzothiazinones (BTZ043) Benzothiazole (TCA1) 4-aminoquinolone piperidine amides 2-carboxyquinoxalines Oxadiazoles Benzo [b]thiophenes Pyrazolopyridones
RNA polymerase	Responsible for transcription	Rifampicin Rifapentine Rifabutin	
Protein synthase	Protein synthesis	Linezolid (https://www.drugbank.ca/drugs/DB00601)	PNU100480 AZD5847
ATP Synthase	ATP synthesis	Bedaquiline	D-Dethiobiotin
Cytidine triphosphate (CTP) synthetase	Catalysis of amination of uridine triphosphate (UTP) into CTP		Thiophenecarboxamide 4-(pyridine 2-yl) thiazole
Transcription factor (IdeR)	Regulating the intracellular levels of iron		Benzo-thiazol benzene sulfonic acid
Lysine-ε-amino transferase (LAT)	Catalysing reversibly the transamination of lysine into α-ketoglutaric acid		Benzothiazole

Table 2.
Mtb *drug targets and the current used drugs [24–26].*

4. Computer resources and tools for tuberculosis drug targets

The development in genomics, coupled with advances in high performance computing and validation of molecular targets, has introduced new approaches to drug discovery that provide a shift from the historical pipeline that focuses on target identification and in most cases involves single targets. In this era of extensive discovery of new chemical entities for treatment of TB and other infectious diseases like HIV/AIDs, a number of research institutes as well as pharmaceutical companies are eagerly developing computational tools and protocols to facilitate drug discovery and development [28]. Genomics provide DNA, RNA, transcriptomic and proteomic data that is housed in a variety of databases and provide resources e.g. from the European Bioinformatics Institute (EBI) https://www.ebi.ac.uk/, and the National Centre for Biotechnology Information (NCBI) https://www.ncbi.nlm.nih.gov/, which can be easily retrieved and analysed, thereby shifting the drug discovery focus from a single to a multi-protein target approach. In this approach *Mtb* genomic data are analysed for network, structure and function of a number of essential proteins that are druggable and validated as potential targets for a number of bactericidal or bacteriostatic chemical compounds. In this section, different databases, resources and tools for target deorphaning are discussed with a particular focus on *Mtb* targets.

The revolution in genomics led to the availability of a number of mycobacterial genomes and the development of a variety of databases consisting of *Mtb* genomic and transcriptomic data. The genomic databases provide information about the structure, function and evolution of *Mtb* genes, whilst the transcriptomics provide information crucial for analysis of gene expression using large scale RNA sequences [29]. On the other hand proteomics provides information about the function, networks and structure of proteins. In their paper, Machado et al. [29] give a detailed summary of most computational resources for TB and we encourage readers to consult the article for more information. Similarly a number of chemogenomic resources and database containing data for *Mtb* ligand annotated targets have been developed. Examples of such databases include the ChEMBL database [30], a database of small bioactive molecules and their targets, TIBLE [31] a database containing MIC and target data for mycobacterial species and TDR targets containing target-ligand information for neglected tropical diseases including TB. The databases are freely available and provide easy access to target-ligand data for *Mtb*. In these databases each target is associated to ligand(s) obtained from bioassays and *vice versa*.

5. Computational target deorphaning techniques

A number of computational methods are being explored in order to identify ligands for both host and pathogen targets and for targets from other organisms like *Plasmodium falciparum* [32]. In most cases two or more complementary ligand-based and structure-based deorphaning approaches are used; statistical methods involving machine learning [8] and deep learning strategies are applied in conjunction with biological and/or biophysical methods to validate the computational results or the computational methods are used to provide the protein-ligand binding information in the absence of X-ray co-crystallised structures of the ligand [12, 13]. In their work, Mendes and Blundell [13] applied cheminformatics to complement current efforts for target identification of fragment-sized molecules that target e.g. the PanC that synthesises pantothenate important for generation of the *Mtb* co-enzyme A. This has led to the identification of 'hotspots' in the binding pockets of a number of proteins, which highlight the most favoured binding spots for the protein. Hotspots and druggability will be discussed in detail in Section 6.

5.1 Ligand-based and structure-based virtual screening methods

Structure-based virtual screening is an approach used in drug discovery to computationally screen small molecule databases for compounds that target proteins of known 3D structure that are experimentally validated. Brain Shoichet [33] has pointed out that this approach was first published in the 1970s, however most new ligands and their targets were not identified until the early 2000. The method offers the opportunity to access a large number of potential new chemical ligands for old and new targets. In the presence of available ligands for named biological targets, ligand-based virtual screening may be used using a variety of techniques ranging from molecular similarity, pharmacophoric search, to machine learning and most recently deep learning.

5.1.1 Structure-based techniques

Structure-based virtual screening plays a significant role in drug discovery in that it is used to identify ligands for biological targets when the 3D structures of the *Mtb* targets from X-ray crystallography, nuclear magnetic resonance (NMR) or cryoelectron microscopy are available in the Protein Data Bank, or homology models available in the CHOPIN database and/or generated in house. This method applies structural data of proteins/receptors to provide small molecules with specific structural attributes for good binding affinity [34]. Generally, the process involves three crucial steps, namely preparation of 3D crystal structures of proteins obtained from the Protein Data Bank (PDB) and the ligand structures, docking calculation and data analysis. Protein structure preparation involves adding hydrogen atoms that are normally missing in the coordinate files, adding missing residues, optimising hydrogen bonds, removing atomic clashes, as well as sampling the degrees of freedom such as flip that are not clear in standard resolution crystal structures, for example the 180° flips of chain terminal rotatable side-chain groups e.g. in shape-symmetric amino acids Asn and Gln, tautomer and/or ionisation state and relaxation of the target and ligand structure [35]. Most docking software is associated with protein and ligand preparation tools, for example Autodock4 or VINA require structures prepared using AutoDockTools (ADT) and the protein preparation script to generate Autodock-type atoms containing Gasteiger charges, and produce the pdbqt files that are compatible with the tool [36]. Similarly, the Primex and Ligprep tools are used to prepare the protein and ligand structures respectively before docking with GLIDE [37]. The quality of input structure files contribute to the quality of the docking results, and the importance of protein and ligand preparation have been highlighted by Sastry [35].

5.1.1.1 Molecular docking

Molecular docking calculations are capable of predicting the binding conformation of ligands inside the binding pocket of a target, as such they are used to map small molecules onto targets and hence provide essential binding information for structure-based drug design. To achieve this, a number of docking algorithms like Autodock [36], perform a stochastic conformational search or e.g. in GLIDE, a [37] that perform a systematic search [34]. In a stochastic search structural parameters, such as torsional, translational and rotational degrees of freedom of the ligand, are randomly modified to generate an ensemble of molecular conformations and increase the chances of finding the energy global minimum, whilst in a systematic conformational search structural features are gradually changed until a local or global minimum is reached [34]. During the search, conformations of a number of potential binding compounds are explored and evaluated using a specific scoring

function. In addition, the conformations are ranked based on their calculated binding energy. Highly ranked compounds are selected as ligands for the target. On the other hand, reverse or inverse docking is used for identifying targets of drug phenotypic hits from a sea of targets. In this way, structure-based screening helps to identify and explain polypharmacology, molecular mechanism of action of substances, facilitate drug repurposing, detect adverse drug reactions and hence toxicity.

5.1.1.2 Deorphaning the HTH transcription regulator, EthR

In an effort to de-orphan the HTH transcription regulator, EthR, and identify the binding mode of the ligand, we docked 200 fragment-like compounds from the Maybridge database to the highest quality crystal structure of the 23 PDB entries using the GOLD algorithm (unpublished work). We used Arpeggio [38], an online tool that identifies non-covalent interactions in protein-structures, to assess the role of each EthR binding site residue and each small-molecule ligand moiety in contributing to protein-ligand interactions. Visual assessment of interactions involved calculating interactions using the Arpeggio web server (http://structure.bioc.cam.ac.uk/arpeggio) and downloading the results as PyMOL session files, to analyse the non-covalent interactions of each residue. We found that in addition to using polar contacts, most ligands are stabilised by a cascade of pi-interactions starting from Tyr103 close to the entrance of the allosteric pocket to Phe114 located close to the HTH-domain and beyond (**Figure 2**). Furthermore, potential ligands for the protein were identified. Information obtained from these results is vital identify ligands with a higher probability of binding to EthR, and so improve the potency and safety of ethionamide (ETH).

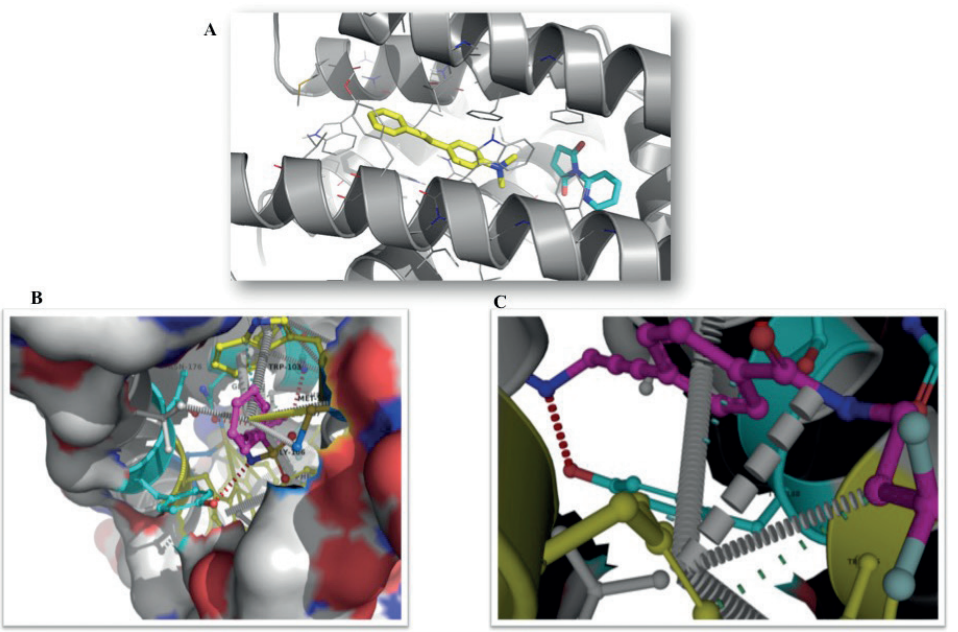

Figure 2.
(A) Binding modes of two fragment-like molecules inside the long cylindrical allosteric binding pocket of EthR defined by five helices. Yellow sticks depict the molecule occupying the upper binding site close to the entrance of the pocket and cyan sticks represent a molecule occupying the inner binding site close to the HTH domain. (B) EthR-ligand interactions involving Trp103 (yellow) at the entrance of the binding pocket of the protein. Ligand atoms and bonds are in pink, grey rings are hydrophobic interactions, red rings show hydrogen bonds. (C) EthR-ligand (pink) interactions involving Phe110 located at the center of the binding pocket of EthR.

Similarly, docking calculations were used to assess binding of ligands identified from for a novel TB drug target, inosine monophosphate dehydrogenase (IMPDH) protein Guab2 that is responsible for the synthesis of xanthosine monophosphate (XMP) from IMP, identified from high throughput screening [12]. Hit compounds were identified in a single shot high-throughput screen, validated by dose response and subjected to further biochemical analysis. The compounds were also assessed using molecular docking experiments, providing a platform for their further optimisation using medicinal chemistry. From the results, it was observed that occupation of the nicotinamide sub-site was correlated with interactions of the ligands with the purine ring of IMP.

5.1.1.3 Applying concerted computational and experimental approaches

Likewise, we used a combination of ligand-based and structure-based chemogenomic approaches, followed by biophysical and biochemical methods, to identify targets for *Mtb* phenotypic hits deposited in the ChEMBL database [11]. In this work, EthR and InhA emerged as potential targets for many of the hits, and some of them displayed activity through both targets. From the 35 predicted EthR inhibitors 25 displayed an inhibition of better than 50%, of which eight showed an IC_{50} better than 50 μM against *Mtb* EthR and three were confirmed to be also active against InhA. Further the EthR-ligand complexes were validated using X-ray crystallography in the Blundell laboratory to give new crystal structures which were deposited in the Protein Data Bank. These results provide new lead compounds that could be further developed into highly active ligands of EthR and InhA and enhance treatment of drug-resistant TB.

6. Modelling proteomes for mycobacteria, hotspots and druggability

A comprehensive understanding of the structural proteomes of mycobacteria is essential for novel drug discovery and elucidating the roles of mutations in drug resistance. Most researchers begin by defining the 3D-structure using X-ray crystallography, NMR or increasingly cryo-EM. For phenotypic screening and understanding off-target hits, where the target is not identified, prior knowledge of the structures of all gene products in the target organism is helpful. This has stimulated the establishment of several consortia in what is usually known as structural genomics, but might more appropriately termed "structural proteomics".

6.1 Evolution of structural genomics consortia

The Structural Genomics Consortium (SGC) [39] which has focused on proteins of interest to medicine, has impressive achievements, in 2011 defining ~40% of the structures of proteins from human parasites deposited in the PDB [40]. The Tuberculosis Structural Genomics Consortium (TBSGC), an international collaboration involving 53 countries, has focused on 3D structures of *Mtb* [40]. This activity and others working on *Mtb* proteomes have deposited 2274 structures in the PDB, but still representing less than 583 gene products, only 13.97% of genome. Although this is a small percentage, it compares impressively with knowledge of protein structures of two other mycobacterial pathogens where there is great clinical interest: for *M. leprae* causing leprosy there are experimentally-defined 3D structures for 15 gene products and for *M. abscessus*, a free living *Mycobacterium*, which is a growing challenge for cystic fibrosis patients, there are 53 experimentally-defined 3D structures in the PDB.

6.2 Comparative 3D modelling of proteins

Comparative modelling proteins, based on the fold recognition and structural alignment with the closest homologues that have experimentally solved structures, began using interactive graphics in the 1970s [41–43]. The development of automated modelling software began in the 1980s, initially with Composer [44] and later developed with Comparer [45] and Modeller [46], based on satisfaction of 3D restraints derived from structurally aligned homologues. Modeller has now been cited ~10,500 times in the literature!

6.2.1 Computational modelling pipelines and structural proteome databases

Rapid progress in this and other related software coupled with increasing computing power has enabled genome scale prediction of protein structures, as a viable alternative to experimental determination. In order to construct computational models of all gene products, which we here refer to as the structural proteome, we identify templates by a sequence-structure homology search using Fugue [47], which uses local-structural-environment-specific substitution tables to predict the likelihood of a common 3D structure. We have incorporated Fugue into a pipeline (Vivace), in which templates are selected from TOCCATA (Ochoa Montaño and Blundell, unpublished), a database of consensus profiles built from CATH 3.5 [48] and SCOP 1.75A [49] based classification of proteins structures (PDB files). PDBs within each profile are clustered based on sequence similarity using CD-HIT [50] and structures are aligned using BATON, a modified version of COMPARER [45]. After further optimization of the clusters by discarding templates with more than 20% difference in sequence identity to the maximum hit, remaining templates are classified into states based on ligand binding and oligomerization. Five different states, known as "liganded-monomeric," "liganded-complexed," "apo-monomeric," "apo-complexed" and "any," are generated in each profile hit. Models are built in each of these states using Modeller 9.10 [46] and refined. Later NDOPE, GA341 [51] Molprobity [52] and SSAG [53] are used to determine the quality of the models.

6.2.2 Mycobacterial proteome databases

The first application of this approach was to construct the Chopin Database (http://mordred.bioc.cam.ac.uk/chopin/about), a database of protein structures for H37Rv strain of *Mtb*. This has provided structures that are reasonably certain for around 65% of gene products. These have proved reliable indicators of the overall structures but may have some uncertainties especially in loop regions and domain-domain relationships. A further ~19% probably have correct folds while the remaining would unlikely to be correct. Nevertheless, compared to those structures defined experimentally by X-ray analysis, this represents a 6-fold increase of structural information available that might be useful in assessing druggability and the impacts of mutations.

Similar models of the structural proteome for *M. abscessus* (Skwark et al., unpublished) and *M. leprae* (Vedithi et al., unpublished) have been developed in the group. In *M. leprae,* of the 1615 gene products, templates were identified for 1429 gene products and we were able to model 1161 proteins with high confidence. A total of 36,408 models were built in different ligand bound and oligomeric states for the 1161 proteins. The distribution of Fugue Z score across models indicates that only 4% of the proteome has no hits and 15% has poor scores. ~80% of the proteome has acceptable and good hits, and the corresponding Z scores. Around 47% of the protein queries identified templates with identity and coverage greater than 40 and 67% of the models in the proteome are of best quality as estimated by NDOPE, GA341, Molprobity and Secondary Structure Agreement (SSAG).

6.2.3 Oligomeric protein models

Current work on structural proteomes includes efforts to extend the modelling pipeline to homo-oligomeric (and eventually hetero-oligomeric) structures using comparative approaches (Malhotra et al., unpublished), extending models and improving models of small molecule complexes, and linking individual protein structures into the metabolic networks and interactions in the cell (Bannerman et al., unpublished). An example of an oligomeric structure is CTP-synthase, encoded by *PyrG*, which is an essential gene in *Mtb* identified by transposon saturation mutagenesis [54] and catalyses ATP-dependent amination of UTP to CTP with either L-glutamine or ammonia. The allosteric effector GTP functions by stabilising the protein conformation that binds to the tetrahedral intermediates formed during glutamine hydrolysis. Its closest homologue in *M. leprae* ML1363 is a target of choice and was modelled using *Vivace* during the proteome modelling exercise. We modelled the apomeric and ligand bound states of the model and oligomerized the protomer using our inhouse oligomerization pipeline. The protomeric and oligomeric states are depicted in **Figure 3A** and **B**.

The models were built by using templates PDB-IDs: 4zdI and 4zdK for PyrG of *Mtb* [55]. Both the templates are 89% identical and 100% coverage to the query sequence. The superposition of the models with the templates indicated a root mean square deviation (RMSD) of 0.758.

6.3 Structural implications of mutations

We have also spent time over 2 decades analysing the impacts of mutations evident in the increasing wealth of available genome sequences for pathogenic mycobacteria and cancers. We originally developed SDM [56] in 1997, a method depending on statistical analysis of environment-dependent amino-acid substitution tables [57, 58]. In 2013 machine learning was introduced with the arrival of Douglas Pires in Cambridge, developing first mCSM for stability [59] followed by several "flavours" including mCSM-PPI for impacts on protein-protein interactions, mCSM-NA [60] for nucleic acid interactions and mCSM-lig for impacts on small-molecule ligand interactions useful for understanding drug resistance [61]. A critical part of using machine learning is to have an extensive database of experimentally-defined impacts of mutations on stability and interactions, such as Platinum by David Ascher when in Cambridge [62], a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes that was developed for mCSM-lig. These two structural approaches to predicting the impacts of mutations (SDM & mCSM) have proved complementary and more reliable than most sequence-only

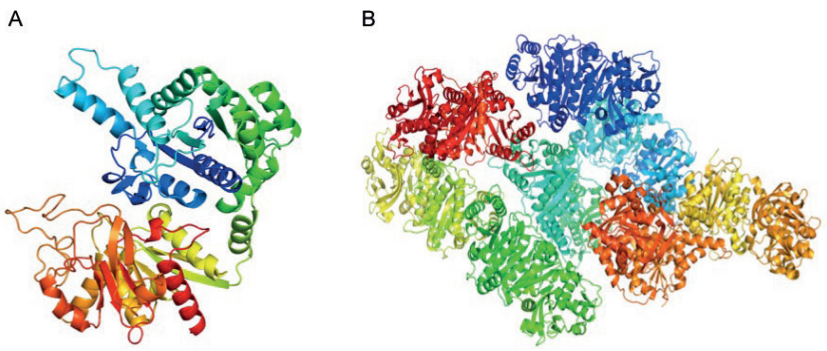

Figure 3.
(A) Protomeric model of PyrG (CTP-Synthase) of M. leprae *modelled with a quality of 4.25 (best).*
(B) Homo-8-mer of PyrG of M. leprae *modelled with a quality of 4.25 (best).*

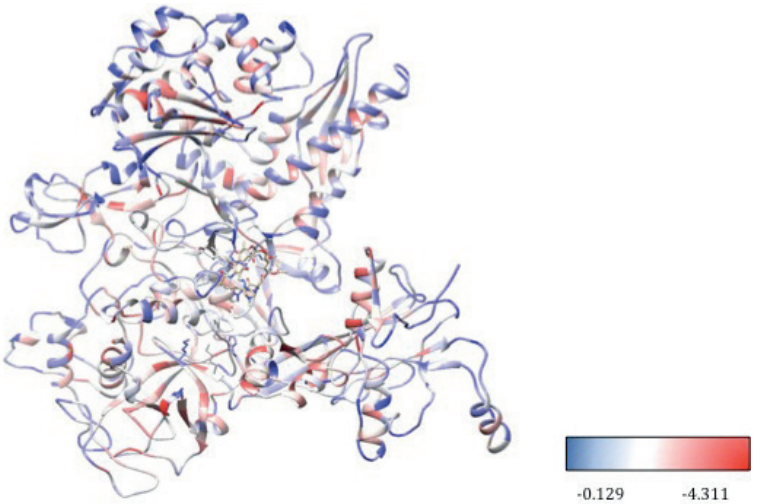

Figure 4.
Indicates the maximum destabilising effect a mutation can induce on the stability of RNA-polymerase β-subunit of M. leprae *(target for rifampin) measured by mCSM-stability.*

methods. They also allow the application of saturation mutagenesis, facilitating *in silico* systematic analysis of mutations [63], an approach now being adopted to whole proteomes where every residue in each of the proteins in the proteome is mutated to all the other 19 amino acids and the effects of the mutations are measured using various methods mentioned above. In structure-guided fragment-based drug discovery, this provides comprehensive information on the regions of the protein that are less likely to lead to drug resistance and therefore can be probed by elaboration of fragments/small molecules. We performed saturation mutagenesis on the drug targets in *M. leprae* for leprosy and the average or highest impact a mutation can induce in each residue position is depicted on the structure (**Figure 4**).

6.4 Active sites, cavities and fragment hotspot maps

Although comparative modelling of homologues in complex with ligands can often give clues about active sites, cofactor binding and substrate or other ligand binding sites, this is not always possible. In order to indicate putative binding sites in the absence of appropriate experimental data, we have exploited cavity-defining software such as VolSite [64] for novel binding site description together with an alignment and comparison tool (Shaper) [65]. We have used FuzCav, a novel alignment-free high-throughput algorithm to compute pairwise similarities between protein-ligand binding sites [66] and GHECOM [67], to study the small pockets that often characterise protein-protein and protein-peptide interactions.

Further to the identification of cavities and pockets, it is also useful to be able to identify hotspots, region(s) of the binding site defined as a major contributor to the binding free energy, and often characterised by their ability to bind fragment-sized organic molecules in well-defined orientations. The usual understanding is that the fragment, with a mixed polar and hydrophobic character, can displace an "unhappy water." We have tried to mimic this *in silico* by using SuperStar [68] to generate atomic interaction propensities on a grid. We then carry out a search with three fragments, each having a six-membered carbon ring, but having a donor, acceptor or a non-polar substituent. The resulting map is convoluted with an estimate of the depth below the surface, which generally appears to correlate with favourable entropic gain on water release on binding of a ligand [69]. The hotspot maps, computed in this way and

indicating donor, acceptor and lipophilic interactions correlate well with experimental binding sites of fragments that can be elaborated in fragment-based discovery. For the ligand bound structures, lower contouring can provide "warm spots" for the binding sites, indicating possibilities for elaborating the fragment in the binding pocket.

The models of individual molecules of the modelled proteome can be individually decorated with the hotspot maps. They give a good indication of the known functional sites on experimentally defined structures of proteins, often demonstrating that a functional site comprises several hotspots involved in binding substrates and cofactors. They also provide a good indication of the location of allosteric sites [70].

7. Conclusion

In summary we can move from the study of individual targets to an understanding of the majority of targets coded by the genome. Indeed, we can build 3D structures for a majority of the genes, so providing a model of the "structural proteome". Hotspots and cavities provide a basis for identification of the ligandability of putative binding sites and have been used in our group to predict pharmacophores that can be used in docking and virtual screening and so deorphaning of mycobacterial proteins.

To identify druggable proteins from the structural proteome, we have adopted a hierarchal selection process wherein chokepoint analysis is initially performed to identify metabolic reactions that are critical to cell survival. Gene products identified in this screen are later subjected to essentiality analysis using either flux balance analysis (FBA) based models or by data from the transposon saturation mutagenesis experiments in the literature. Genes that are essential are chosen at this stage and understanding of the gene expression profiles in different growth conditions is analysed. Genes whose expression is condition specific are excluded. Later for the selected genes, the structural information of the corresponding proteins is analysed in the context of prior knowledge and attempts in drug discovery, druggable pockets and fragment hotspots maps, small molecule bound states, non-human homologue, non-homologous to human microbiome, cellular localization and biochemical properties of the proteins. Structure-guided virtual screening is performed on the selected drug targets with a choice of fragment and compound libraries using CCDC Gold (The Cambridge Crystallographic Data Centre) [71]. Best poses with good scores lead the experimental process of structure-guided fragment-based drug discovery.

The challenge now is to test the computational methods outlined here for identifying ligands and understanding the druggability of the proteome—several thousand gene products from the whole genome of *Mtb*. We can then begin to assess the degree to which we can de-orphan the many *Mtb* proteins that have until now not featured as targets in the worldwide efforts to combat the global challenge of TB to the health and well-being of human kind.

Acknowledgements

LYB and GCM are grateful to Chinhoyi University of Technology for their support in introducing computational drug discovery and development research work at the University and all our collaborators. TLB and SCV thank the Gates Foundation, the Cystic Fibrosis Trust and the American Leprosy Mission for their funding of computational and experimental work on approaches to combating disease from mycobacterial infections. They also thank colleagues in Cambridge and elsewhere who have contributed over the years to our efforts to develop new approaches to structural biology, computational bioinformatics and drug discovery.

Author details

Lorraine Yamurai Bishi[1], Sundeep Chaitanya Vedithi[2], Tom L. Blundell[2] and Grace Chitima Mugumbate[1*]

1 Department of Chemistry, School of Natural Sciences and Mathematics, Chinhoyi University of Technology, Chinhoyi, Zimbabwe

2 Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

*Address all correspondence to: gmugumbate@cut.ac.zw

IntechOpen

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Nguta JM, Appiah-Opong R, Nyarko AK, Yeboah-Manu D, Addo PGA. Medicinal plants used to treat TB in Ghana. International Journal of Mycobacteriology. 2015;**4**:116-123. DOI: 10.1016/j.ijmyco.2015.02.003

[2] Jiang J, Gu J, Zhang L, Zhang C, Deng X, Dou T, et al. Comparing *Mycobacterium tuberculosis* genomes using genome topology networks. BMC Genomics. 2015;**16**:1-10. DOI: 10.1186/s12864-015-1259-0

[3] Shah I. Drug Resistant Tuberculosis Children in India. Pediatric Oncall Journal. 2012;**9**(5). DOI: 10.7199/ped.oncall.2012.27

[4] Nguta JM, Appiah-Opong R, Nyarko AK, Yeboah-Manu D, Addo PGA. Current perspectives in drug discovery against tuberculosis from natural products. International Journal of Mycobacteriology. 2015;**4**:165-183. DOI: 10.1016/j.ijmyco.2015.05.004

[5] Janbaz KH, Qadir MI, Ahmad B, Sarwar A, Yaqoob N, Masood MI. Tuberculosis burning issues: Multidrug resistance and HIV-coinfection. Critical Reviews in Microbiology. 2012;**38**:267-275. DOI: 10.3109/1040841X.2012.664539

[6] Anishetty S, Pulimi M, Pennathur G. Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. Computational Biology and Chemistry. 2005;**29**:368-378. DOI: 10.1016/j.compbiolchem.2005.07.001

[7] Hart CP. Finding the target after screening the phenotype. Drug Discovery Today. 2005;**10**:513-519. DOI: 10.1016/S1359-6446(05)03415-X

[8] Mugumbate G, Abrahams KA, Cox JAG, Papadatos G, van Westen G, Lelièvre J, et al. Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and in vitro validation. PLoS One. 2015;**10**:e0121492. DOI: 10.1371/journal.pone.0121492

[9] Bajorath J. Computer-aided drug discovery. F1000 Research. 2015;**4**:630. DOI: 10.12688/f1000research.6653.1

[10] Bender A, Young D, Jenkins J, Serrano M, Mikhailov D, Clemons P, et al. Chemogenomic data analysis: Prediction of small-molecule targets and the advent of biological fingerprints. Combinatorial Chemistry & High Throughput Screening. 2007;**10**:719-731. DOI: 10.2174/138620707782507313

[11] Mugumbate G, Mendes V, Blaszczyk M, Sabbah M, Papadatos G, Lelievre J, et al. Target identification of *Mycobacterium tuberculosis* phenotypic hits using a concerted chemogenomic, biophysical, and structural approach. Frontiers in Pharmacology. 2017;**8**:681. DOI: 10.3389/fphar.2017.00681

[12] Cox JAG, Mugumbate G, Del Peral LVG, Jankute M, Abrahams KA, Jervis P, et al. Novel inhibitors of *Mycobacterium tuberculosis* GuaB2 identified by a target based high-throughput phenotypic screen. Scientific Reports. 2016;**6**:1-10. DOI: 10.1038/srep38986

[13] Mendes V, Blundell TL. Targeting tuberculosis using drug design. Drug Discovery Today. 2016;**00**:1-9. DOI: 10.1016/j.drudis.2016.10.003

[14] Cole RBST, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;**393**:537-544. DOI: 10.1038/31159

[15] Smith I. *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. Clinical

Microbiology Reviews. 2003;**16**:463-496. DOI: 10.1128/CMR.16.3.463

[16] Hoagland DT, Liu J, Lee RB, Lee RE. New agents for the treatment of drug-resistant *Mycobacterium tuberculosis* ☆. Advanced Drug Delivery Reviews. 2016;**102**:55-72. DOI: 10.1016/j.addr.2016.04.026

[17] World Health Organisation. Global tuberculosis Report 2017. Geneva; 2017

[18] Zhang Y, Post-Martens K, Denkin S. New drug candidates and therapeutic targets for tuberculosis therapy. Drug Discovery Today. 2006;**11**:21-27. DOI: 10.1016/S1359-6446(05)03626-3

[19] Pawlowski A, Jansson M, Sköld M, Rottenberg ME, Källenius G. Tuberculosis and HIV co-infection. PLoS Pathogens. 2012;**8**(2):e1002464. https://doi.org/10.1371/journal.ppat.1002464

[20] Metcalfe JZ, Porco TC, Westenhouse J, Damesyn M, Facer M, Hill J, et al. Tuberculosis and HIV co-infection, California, USA, 1993-2008. Emerging Infectious Diseases. 2013;**19**:400-406. DOI: 10.3201/eid1903.121521

[21] Worley MV, Estrada SJ. Bedaquiline: A novel antitubercular agent for the treatment of multidrug-resistant tuberculosis. The Journal of Human Pharmacology and Drug Therapy. 2014;**34**(11):1187-1197

[22] Gawad J, Bonde C. Current affairs, future perspectives of tuberculosis and antitubercular agents. The Indian Journal of Tuberculosis. 2018;**65**:15-22. DOI: 10.1016/j.ijtb.2017.08.011

[23] Lamichhane G. Novel targets in M. tuberculosis: Search for new drugs. Trends in Molecular Medicine. 2011;**17**:25-33. DOI: 10.1016/j.molmed.2010.10.004

[24] Lewis K. Platforms for antibiotic discovery. Nature Reviews. Drug Discovery. 2013;**12**:371-387. DOI: 10.1038/nrd3975

[25] Kaneko T, Cooper C, Mdluli K. Challenges and opportunities in developing novel drugs for TB. Future Medicinal Chemistry. 2011;**3**:1373-1400. DOI: 10.4155/fmc.11.115

[26] Campaniço A, Moreira R, Lopes F. Drug discovery in tuberculosis. New drug targets and antimycobacterial agents. European Journal of Medicinal Chemistry. 2018;**150**:525-545. DOI: 10.1016/j.ejmech.2018.03.020

[27] Chiliza TE, Pillay M, Pillay B. Identification of unique essential proteins from a *Mycobacterium tuberculosis* F15/LAM4/KZN phage secretome library. Pathogens and Disease. 2017;**75**:1-10. DOI: 10.1093/femspd/ftx001

[28] Kapetanovic IM. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. Chemico-Biological Interactions. 2008;**171**:165-176. DOI: 10.1016/j.cbi.2006.12.006

[29] Machado E, Cerdeira C, de Miranda AB, Catanho M. Web resources on tuberculosis: Information, research, and data analysis. In: Mycobacterium-research and development. IntechOpen

[30] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. {ChEMBL}: A large-scale bioactivity database for drug discovery. Nucleic Acids Research. 2012;**40**:D1100-D1107. DOI: 10.1093/nar/gkr777

[31] Malhotra S, Mugumbate G, Blundell TL, Higueruelo AP. TIBLE: A web-based, freely accessible resource for small-molecule binding data for mycobacterial species. Database (Oxford). 2017;**2017**:1-7. DOI: 10.1093/database/bax041

[32] Mugumbate G, Newton AS, Rosenthal PJ, Gut J, Moreira R, Chibale K, et al.

Novel anti-plasmodial hits identified by virtual screening of the ZINC database. Journal of Computer-Aided Molecular Design. 2013;**27**:859-871. DOI: 10.1007/s10822-013-9685-z

[33] Shoichet BK. Virtual screening of chemical libraries. Nature. 2004;**432**:862-865. DOI: 10.1038/nature03197

[34] Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. 2015. DOI: 10.3390/molecules200713384

[35] Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. Journal of Computer-Aided Molecular Design. 2013;**27**:221-234. DOI: 10.1007/s10822-013-9644-8

[36] Morris G, Huey R. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of Computational Chemistry. 2009;**30**:2785-2791. DOI: 10.1002/jcc.21256.AutoDock4

[37] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of Medicinal Chemistry. 2004;**47**:1739-1749. DOI: 10.1021/jm0306430

[38] Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. Journal of Molecular Biology. 2017;**429**:365-371. DOI: 10.1016/j.jmb.2016.12.004

[39] Colwill K, Renewable Protein Binder Working Group, Gräslund S. A roadmap to generate renewable protein binders to the human proteome. Nature Methods. 2011;**8**:551-558. DOI: 10.1038/nmeth.1607

[40] Chim N, Habel JE, Johnston JM, Krieger I, Miallau L, Sankaranarayanan R, et al. The TB structural genomics consortium: A decade of progress. Tuberculosis. 2011;**91**:155-172. DOI: 10.1016/j.tube.2010.11.009

[41] Bedarkar S, Blundell TL, Dockerill S, Tickle IJ, Wood SP. Polypeptide hormone-receptor interactions: The structure and receptor binding of insulin and glucagon. In: Molecular interactions and activity in proteins. Amsterdam, Oxford, New York: Excerpta Medica. 1978;**60**:105

[42] Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature. 1987;**326**:347-352. DOI: 10.1038/326347a0

[43] Blundell T, Sibanda BL, Pearl L. Three-dimensional structure, specificity and catalytic mechanism of renin. Nature. 1983;**304**:273-275. DOI: 10.1038/304273a0

[44] Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Knowledge based modelling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Engineering, Design & Selection. 1987;**1**:377-384. DOI: 10.1093/protein/1.5.377

[45] Šali A, Blundell TL. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. Journal of Molecular Biology. 1990;**212**:403-428. DOI: 10.1016/0022-2836(90)90134-8

[46] Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology. 1993;**234**:779-815. DOI: 10.1006/jmbi.1993.1626

[47] Shi J, Blundell TL, Mizuguchi K. FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. Journal of Molecular Biology. 2001;**310**:243-257. DOI: 10.1006/jmbi.2001.4762

[48] Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J. CATH—A hierarchic classification of protein domain structures. Structure. 1997;**5**:1093-1109. DOI: 10.1016/S0969-2126(97)00260-8

[49] Rost B, Brenner SE, Chothia C, Hubbard TJP, Murzin AG, Li WW, et al. CKAAPs DB: A conserved key amino acid positions database. Methods in Enzymology. 2002;**28**:409-411. DOI: 10.1016/S0076-6879(96)66039-X

[50] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;**28**:3150-3152. DOI: 10.1093/bioinformatics/bts565

[51] Melo F, Sali A. Fold assessment for comparative protein structure modeling. Protein Science. 2007;**16**:2412-2426. DOI: 10.1110/ps.072895107

[52] Davis IW, Murray LW, Richardson JS, Richardson DC. MolProbity: Structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Research. 2004;**32**:W615-W619. DOI: 10.1093/nar/gkh398

[53] Eramian D, Shen M, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Science. 2006;**15**:1653-1666. DOI: 10.1110/ps.062095806

[54] Dejesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. MBio Journal. 2017;**8**:e02133-e02116. DOI: 10.1128/mBio.02133-16

[55] Mori G, Chiarelli LR, Esposito M, Makarov V, Bellinzoni M, Hartkoorn RC, et al. Thiophenecarboxamide derivatives activated by EthA kill *Mycobacterium tuberculosis* by inhibiting the CTP synthetase PyrG. Chemistry & Biology. 2015;**22**:917-927. DOI: 10.1016/j.chembiol.2015.05.016

[56] Topham CM, Srinivasan N, Blundell TL. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Engineering. 1997;**10**:7-21. DOI: 10.1093/protein/10.1.7

[57] Worth CL, Preissner R, Blundell TL. SDM—A server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Research. 2011;**39**:W215-W222. DOI: 10.1093/nar/gkr363

[58] Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: A server for predicting effects of mutations on protein stability. Nucleic Acids Research. 2017;**45**:W229-W235. DOI: 10.1093/nar/gkx439

[59] Pires DEV, Ascher DB, Blundell TL. MCSM: Predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;**30**:335-342. DOI: 10.1093/bioinformatics/btt691

[60] Pires DEV, Ascher DB. MCSM-NA: Predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Research. 2017;**45**:W241-W246. DOI: 10.1093/nar/gkx236

[61] Pires DEV, Blundell TL, Ascher DB. MCSM-lig: Quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Scientific Reports. 2016;**6**:29575. DOI: 10.1038/srep29575

[62] Pires DEV, Blundell TL, Ascher DB. Platinum: A database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Research. 2015;**43**:D387-D391. DOI: 10.1093/nar/gku966

[63] Pires DEV, Chen J, Blundell TL, Ascher DB. In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Scientific Reports. 2016;**6**:19848. DOI: 10.1038/srep19848

[64] Desaphy J, Azdimousa K, Kellenberger E, Rognan D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. Journal of Chemical Information and Modeling. 2012;**52**:2287-2299. DOI: 10.1021/ci300184x

[65] Ehrt C, Brinkjost T, Koch O. Impact of binding site comparisons on medicinal chemistry and rational molecular design. Journal of Medicinal Chemistry. 2016;**59**:4121-4151. DOI: 10.1021/acs.jmedchem.6b00078

[66] Weill N, Rognan D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. Journal of Chemical Information and Modeling. 2010;**50**:123-135. DOI: 10.1021/ci900349y

[67] Kawabata T. Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins: Structure, Function, and Bioinformatics. 2010;**78**:1195-1211. DOI: 10.1002/prot.22639

[68] Verdonk ML, Cole JC, Taylor R. SuperStar: A knowledge-based approach for identifying interaction sites in proteins. Journal of Molecular Biology. 1999;**289**:1093-1108. DOI: 10.1006/jmbi.1999.2809

[69] Radoux CJ, Olsson TSG, Pitt WR, Groom CR, Blundell TL. Identifying interactions that determine fragment binding at protein hotspots. Journal of Medicinal Chemistry. 2016;**59**:4314-4325. DOI: 10.1021/acs.jmedchem.5b01980

[70] Thomas SE, Mendes V, Kim SY, Malhotra S, Ochoa-Montaño B, Blaszczyk M, et al. Structural biology and the design of new therapeutics: From HIV and cancer to mycobacterial infections: A paper dedicated to John Kendrew. Journal of Molecular Biology. 2017;**429**:2677-2693. DOI: 10.1016/j.jmb.2017.06.014

[71] Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. Journal of Molecular Biology. 1997;**267**:727-748. DOI: 10.1006/jmbi.1996.0897