HHS Public Access

Author manuscript

*Tuberculosis (Edinb)*. Author manuscript; available in PMC 2016 March 01.

Published in final edited form as:

*Tuberculosis (Edinb)*. 2015 March ; 95(2): 142–148. doi:10.1016/j.tube.2014.12.003.

# Increasing the Structural Coverage of Tuberculosis Drug Targets

**Loren Baugh**[a,b], **Isabelle Phan**[a,b], **Darren W. Begley**[a,c], **Matthew C. Clifton**[a,c], **Brianna Armour**[a,c], **David M. Dranow**[a,c], **Brandy M. Taylor**[a,c], **Marvin M. Muruthi**[a,c], **Jan Abendroth**[a,c], **James W. Fairman**[c], **David Fox III**[c], **Shellie H. Dieterich**[c], **Bart L. Staker**[a,b], **Anna S. Gardberg**[a,c,d], **Ryan Choi**[a,e], **Stephen N. Hewitt**[a,e], **Alberto J. Napuli**[a,e], **Janette Myers**[a,e], **Lynn K. Barrett**[a,e], **Yang Zhang**[a,b], **Micah Ferrell**[a,b], **Elizabeth Mundt**[a,b], **Katie Thompkins**[a,b], **Ngoc Tran**[a,b], **Sally Lyons-Abbott**[a,b], **Ariel Abramov**[a,b], **Aarthi Sekar**[a,b], **Dmitri Serbzhinskiy**[a,b], **Don Lorimer**[a,c], **Garry W. Buchko**[a,f], **Robin Stacy**[a,b], **Lance J. Stewart**[a,c,g], **Thomas E. Edwards**[a,c], **Wesley C. Van Voorhis**[a,e,h,i], and **Peter J. Myler**[a,b,h,j,*]

[a]Seattle Structural Genomics Center for Infectious Disease

[b]Seattle Biomedical Research Institute, 307 Westlake Ave N, Suite 500, Seattle, Washington 98109

[c]Beryllium, 7869 NE Day Road West, Bainbridge Island, Washington 98110

[d]EMD Serono Research & Development Institute, Inc., 45A Middlesex Turnpike, Billerica, Massachusetts 01821

[e]Department of Medicine, Division of Allergy and Infectious Disease, University of Washington, 750 Republican Street, E-701, Box 358061, Seattle, Washington 98109

[f]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352.

[g]Institute for Protein Design, University of Washington, Box 357350, Seattle, Washington 98195

[h]Department of Global Health, University of Washington, Box 359931, Seattle, Washington, 98195

[i]Department of Microbiology, University of Washington, Box 357735, Seattle, Washington 98195

[j]Department of Biomedical Informatics and Medical Education, University of Washington, Box 358047, Seattle, Washington 98195

## Abstract

High-resolution three-dimensional structures of essential *Mycobacterium tuberculosis* (Mtb) proteins provide templates for TB drug design, but are available for only a small fraction of the Mtb proteome. Here we evaluate an intra-genus “homolog-rescue” strategy to increase the



*Corresponding author: peter.myler@seattlebiomed.org, Seattle Biomedical Research Institute, 307 Westlake Ave N, Suite 500, Seattle, Washington 98109, Phone: 206-256-7332; Fax: 206-256-7229 .

structural information available for TB drug discovery by using mycobacterial homologs with conserved active sites. Of 179 potential TB drug targets selected for x-ray structure determination, only 16 yielded a crystal structure. By adding 1675 homologs from nine other mycobacterial species to the pipeline, structures representing an additional 52 otherwise intractable targets were solved. To determine whether these homolog structures would be useful surrogates in TB drug design, we compared the active sites of 106 pairs of Mtb and non-TB mycobacterial (NTM) enzyme homologs with experimentally determined structures, using three metrics of active site similarity, including superposition of continuous pharmacophoric property distributions. Pair-wise structural comparisons revealed that 19/22 pairs with >55% overall sequence identity had active site Cα RMSD <1Å, >85% side chain identity, and ≥80% $PS_{APF}$ (similarity based on pharmacophoric properties) indicating highly conserved active site shape and chemistry. Applying these results to the 52 NTM structures described above, 41 shared >55% sequence identity with the Mtb target, thus increasing the effective structural coverage of the 179 Mtb targets over three-fold (from 9% to 32%). The utility of these structures in TB drug design can be tested by designing inhibitors using the homolog structure and assaying the cognate Mtb enzyme; a promising test case, Mtb cytidylate kinase, is described. The homolog-rescue strategy evaluated here for TB is also generalizable to drug targets for other diseases.

**Keywords**

Drug discovery; homolog-rescue; structural genomics; enzyme active site

## 1. Introduction

One strategy for discovering new TB drugs is to use whole-cell screens of molecular libraries to identify compounds that kill or slow the growth of *Mycobacterium tuberculosis* (Mtb), the causative agent of TB (1). Follow-up studies on these inhibitors can rule out non-specific toxicity, establish pharmacokinetic/dynamic properties, and identify target proteins in the Mtb organism. Ideally, a new library consisting of variants of the lead compounds is then designed to improve the steric and chemical match with the active site of the Mtb targets, and these molecules are subsequently screened for enhanced activity against both the targets and whole cells. This cycle can be repeated until molecules with sufficiently high binding affinity and potency are identified (2). Such an approach to drug discovery is enhanced by a high-resolution three-dimensional structure of the drug target to serve as a template against which inhibitors and inhibitor libraries can be refined. However, only ~10% of the Mtb proteome has been structurally characterized, representing a large blind spot for TB drug development. One reason for low proteomic coverage is that obtaining x-ray crystal structures remains challenging despite technological improvements in gene-to-structure pipelines (3, 4). Cloned genes often fail to express proteins that are soluble or crystallizable, and even when crystals are obtained they sometimes do not produce high resolution x-ray diffraction data, resulting in low gene-to-structure success rates (typically <10%) for large-scale efforts.

When structure determination for a desired target fails, one approach is to engineer genetic variants containing terminal additions or deletions, loop deletions, or point mutations at

crystal contacts, which can produce proteins with improved expression, solubility, and crystallization properties. Combinatorial libraries of such mutants can be used to screen for variants with improved properties (5-8). However, the mutations may also disrupt an active site of the protein, resulting in structures that are less useful in inhibitor design. An alternative approach is to use homologs (proteins descended from a single ancestral gene) from a related species to obtain a surrogate structure for the desired target, since sequence variations between homologs can result in more favorable solubility and crystallization properties (9). Ideally, a homolog with matching active site and substrate specificity would be selected. However, identifying such homologs on a large scale is problematic because most proteins have not been characterized experimentally, and identifying true orthologs requires extensive phylogenetic analysis (10). Instead, targets are more often selected based on sequence similarity, an approach that has been used by several structural genomic projects (4, 9, 11-15).

Homolog structures have proved useful as surrogates in drug discovery when the desired structure was unavailable. The anti-cancer drug Nolatrexed, a 5-substituted quinazolinone, was developed using the structure of thymidylate synthase from *Escherichia coli* (46% overall amino acid sequence identity with the human homolog) (16, 17), and the hypertension drug Captopril was developed by optimizing inhibitors targeting angiotensin-converting enzyme based on a structure of bovine carboxypeptidase A (similar to the human homolog only in its active site) (18). Kinase inhibitors that block transmission of *Plasmodium falciparum* (the causative agent of malaria) to mosquitoes were identified using structures of calcium-dependent protein kinases from *Toxoplasma gondii* and *Cryptosporidium parvum* (74% and 61% sequence identity versus the *P. falciparum* homolog, respectively), because no corresponding *P. falciparum* structure was available (19). Importantly for this study, Bedaquiline (Sirturo), the first approved anti-TB drug with a new mechanism of action in over forty years, was discovered in a whole cell screen of compounds against *Mycobacterium smegmatis*, rather than against *M. tuberculosis*, and targets ATP synthases in both species (proteins sharing 92% sequence identity) (20).

Here, by combining a large-scale effort to obtain x-ray structures for 179 potential TB drug targets with a Protein Data Bank-wide comparison of mycobacterial enzyme active sites, we demonstrate that the coverage of TB drug targets can be increased several-fold with structures having conserved active sites by using homologs selected from within the genus.

## 2. Results

### 2.1 Structure determination

One hundred and seventy-nine Mtb proteins (listed in Table S1) were selected based on their potential value as TB drug targets (see Methods for selection criteria). However, an x-ray structure was solved for only 16 (9%) of the 179 Mtb targets (see Fig. 1). Therefore, an additional 1675 potential homologs were selected from nine other non-TB mycobacterial (NTM) species using a BLASTP search for proteins with >40% sequence similarity (typically equivalent to ~20-25% identity) over >70% of their sequence versus the Mtb protein. This threshold was chosen because proteins with greater than ~40% sequence similarity tend to have similar overall structures (21). This resulted in another 154 structures, including

different ligand-binding states, for one or more homologs of 52 additional Mtb targets (see Table S2). Thus, the overall structural coverage of the Mtb targets was increased from 9% (16/179) to 38% (68/179).

### 2.2 Success rates for different species

The overall success rate for sequence-to-structure was 9.7%, with 145 of the 1501 targets resulting in at least one structure in PDB (Table 1). Three hundred and forty-three targets for which work was stopped when a structure was obtained for a highly similar (>70% sequence identity) protein were excluded from the counts and success rates shown in Table 1. Species-specific overall success rates ("In PDB") were similar to Mtb (11%) in most other species (7-13%), but were lower for *M. ulcerans* (3%), *M. leprae* (3%) and *M. bovis* (0%), although the number of targets selected was low for the last species. The overall step-wise success rates (far right column) varied substantially; being lowest for "HQ data" (42%) and "Soluble" (57%), and highest for "Cloned" (90%) and "In PDB" (88%). Species-specific step-wise success rates can be obtained by dividing the cumulative success rate at one step by that at the previous step. *M. smegmatis* showed a significantly higher rate (72%) of soluble protein expression (given successful cloning) than Mtb (60%), while the success rates for *M. bovis* (33%), *M. leprae* (30%), and *M. ulcerans* (35%) were all significantly lower. Most species showed similar success rates for protein purification and crystallization (combined), with only *M. smegmatis* (60%) being significantly higher than Mtb (44%). *M. paratuberculosis* (29%) and *M. bovis* (0%) had significantly lower success rates for obtaining diffraction data from crystals, with all other species enjoying similar success rates to Mtb (56%), except for *M. ulcerans* (32%), although the latter was not statistically significant.

### 2.3 Enzyme structure and active site comparisons

Are these NTM structures likely to be useful for TB drug design? To address this question, we examined the structures solved in this study, as well as those in the RCSB Protein Data Bank (PDB) (22) and identified 106 pairs of Mtb and NTM enzymes with >25% sequence identity, a known active site in the Mtb enzyme (based on a substrate-bound structure), and x-ray structures available for the same ligand binding state (*i.e.* same substrate or large co-factor bound, or no large ligand bound) (Fig. 2; Table S3). Thirty of the 106 pairs include a structure solved in this study.

We initially compared the overall and active site backbone structures of enzyme pairs by calculating their Cα root-mean-square deviation (RMSD), a measure of the average distance between backbone atoms in two aligned structures. As shown in Fig. 3*A*, the 106 enzyme pairs fell into two distinct groups: 22 pairs with >55% sequence identity and 84 with <42% sequence identity. All of the former showed Cα RMSDs <1Å, indicating similar overall structures, while the majority (60/84) of the latter have Cα RMSDs >1Å, indicating more dissimilar overall structures. This relationship between overall structural similarity and protein global sequence identity has been described previously (23, 24). Importantly, almost all (21/22) pairs with >55% sequence identity represent members of the same OrthoMCL protein family (25), while most (72/84) pairs with <42% sequence identity belong to different families. When the structure comparison was restricted to active site residues

(defined as those within 4Å of the substrate in a substrate-bound structure), 20/22 enzyme pairs with >55% overall sequence identity have an active site Cα RMSD <1Å, while most (43/84) pairs with <42% sequence identity have an active site Cα RMSD >1Å (Fig. 3*B*).

To further compare active sites, we used two additional metrics. The first was active site amino acid composition, measured as the fraction of active site side chains that are identical following sequence and structure alignment (Fig. 4*A*). All 22 pairs with >55% global sequence identity have >80% (often 100%) identity in active site side chains; while 80/84 of pairs with <42% global sequence identity have <80% identical active site side chains. The second was optimized superposition of continuous pharmacophoric property distributions, a measure of similarity of active site shape and chemistry (Fig. 4*B*). Active site x-ray structures were converted into atomic property fields representing seven pharmacophoric or chemical properties, optimally superimposed by a Monte Carlo procedure, and scored (26). The $PS_{APF}$ (Pocket Similarity based on Atomic Property Field) scores represent fractional similarity: *i.e.* if one site has half of the atoms missing, but is otherwise identical to the other, the score would be 50%. Based on this metric, 19/22 pairs with >55% global sequence identity have $PS_{APF}$ scores ≥80%, indicating a highly similar active site shape and chemistry, while 77/84 pairs with <42% sequence identity have $PS_{APF}$ <80%, suggesting a less similar active site architecture.

Active site similarity can also be measured using the RMSD over all shared atoms in the active sites. Since unpaired atoms are excluded from all-atom RMSD measurements when side chains differ, these measurements are most useful for comparing active sites with identical or nearly identical side chains. We performed all-atom RMSD measurements for all structure pairs with at least 90% active site side chain identity (Table S3), and plotted the results versus $PS_{APF}$ scores (Fig. S1). The results show a strong linear correlation (*R* value −0.90) between all-atom RMSD and $PS_{APF}$ values. Furthermore, at $PS_{APF}$ >80%, in 14/16 cases the RMSD value is <1Å, indicating nearly identical side chain orientations. For comparing active sites with lower side chain identity, $PS_{APF}$ measurements have the advantage of including all functional groups in the comparison.

Active site superpositions for four enzyme pairs are shown in Fig. 5 and active site superpositions for all 33 enzyme pairs in the same OrthoMCL family can be found in Dataset S1. In Fig. 5*A*-*C*, the active sites of both enzymes in each pair are extremely similar ($PS_{APF}$ >93%), even though global sequence identity ranges from 63% to 95%. In Fig. 5*D*, the active sites are less similar ($PS_{APF}$ = 71%) despite high enzyme sequence identity (82%). In most cases for which $PS_{APF}$ >80%, the active sites have nearly identical side chain orientations.

Based on these three metrics of active site similarity, mycobacterial enzymes with >55% sequence identity are extremely likely to share similar active sites, since 19/22 such pairs in the data set showed <1Å active site Cα RMSD, >85% active site side chain identity, and ≥80% $PS_{APF}$. Thus, our x-ray structure determination effort increased the active site structural coverage of 179 potential TB drug targets (Mtb proteins) from 16 (9%) to 57 (32%) by using NTM homologs with >55% global sequence identity.

## 3. Discussion

When a structure for a desired drug target is unavailable, the ideal surrogate would possess a (nearly) identical active site or binding pocket; such that a compound designed using the surrogate will also inhibit the desired target. However, the best candidates to obtain such a structure - an ortholog or homolog with shared substrate specificity - are often unknown, since for most proteins functional annotation is based solely on sequence similarity. An alternate basis for selection is amino acid sequence identity, since proteins with high sequence identity tend to share similar structures (23). Prior to this study, the relationship between protein active site similarity and global sequence identity was unclear. The coincidence of similar pockets in proteins with comparable overall structures and sequences has been noted (27). Sequence comparison and enzyme functional annotation have been used to show that above 60% sequence identity, enzyme function conservation can be inferred to all four Enzyme Commission (EC) digits with 90% accuracy (28). Since the last EC digit specifies substrate specificity, proteins with >60% sequence identity may be expected to have similar substrate-binding pockets.

Here, using 106 pairs of Mtb and NTM enzyme structures, and three metrics of active site similarity, we find that enzyme active site shape and chemistry is highly conserved in pairs with >55% global sequence identity. Active site side chains often change conformation upon ligand binding (29), but such differences were avoided by comparing structures in the same ligand binding state (unbound, or same ligand or large cofactor bound). There may also be exceptions where only a few amino acid differences between enzymes results in dramatically different active sites (despite high global sequence identity), but such exceptions were not observed in the current dataset. Conversely, most enzyme pairs with lower global sequence identity (<42%), show considerable variation in active site structure, although some may possess very similar active sites (6/84 such pairs in the data set showed highly similar active site shape and chemistry, Fig 4*B*).

Based on high degree of active site similarity between mycobacterial enzymes with >55% global sequence identity, we suggest that such homolog structures will be useful as surrogates in TB drug discovery when the Mtb target structure is unavailable. Obviously, this needs to be tested by designing inhibitors using such homolog structures and assaying them against the cognate Mtb enzymes. One promising test case using materials available from this study is Mtb cytidylate kinase, which plays an important role in the synthesis and salvage pathways of DNA and RNA precursors (30, 31). The Mtb enzyme is essential *in vitro* (32, 33), and differs in its substrate specificity from the human ortholog, UMP/CMP kinase (34-36). While we were able to purify Mtb cytidylate kinase, the protein did not crystallize, and no structure is available in the PDB. However, structures were solved for homologs from *M. smegmatis* (PDB ID: 3R20) and *M. abscessus* (4DIE). These proteins have 68% and 74% sequence identity to the Mtb cytidylate kinase, respectively, and 73% sequence identity to each other. Their active sites show nearly identical side chain orientations surrounding the substrate (Fig. 6). Based on their high sequence identities versus Mtb cytidylate kinase, it is likely that the active site will also be conserved in the Mtb structure. It will be interesting to see whether an inhibitor designed against these *M.*

*smegmatis* and *M. abscessus* homolog structures has activity against *M. tuberculosis* cytidylate kinase.

## 4. Materials and Methods

### 4.1 Target selection

*M. tuberculosis* H37Rv targets were selected using several selection methods followed by a set of negative filters. Most of the Mtb targets (139/179) were selected internally using a BLASTP search of the Mtb H37Rv genome against the DRUGBANK database (www.drugbank.ca), retaining hits with >50% sequence identity over >75% of the sequence. Other targets were selected using a literature survey to identify orthologs of known drug targets (8/179) or other promising drug candidates (12/179). The remaining 20 targets were Community Requests. Negative filters included the elimination of proteins having >500 amino acids, >8 cysteine residues, and/or containing transmembrane spanning domains. A BLASTP search of the remaining candidate sequences against TARGETDB (http://targetdb-dev.rutgers.edu/targetdb-dev) was performed and those with >95% sequence identity over >80% of the sequence were eliminated. The remaining sequences were BLASTed against each other with a cut-off of >75% similarity over >75% of the sequence to eliminate redundant in-paralogs.

To select the 1675 non-TB mycobacterial (NTM) homologs, the following nine genomes were searched: *M. abscessus* ATCC 19977/DSM 44196, *M. avium* 104, *M. bovis* AF2122/97/ATCC BAA-935, *M. leprae* Br4923, *M. marinum* ATCC BAA-535/M, *M. paratuberculosis* ATCC BAA-968/K-10, *M. smegmatis* ATCC 700084/mc(2) 155, *M. ulcerans* Agy99, and *M. thermoresistibile* ATCC 19527/Tsukamura. Complete sets of protein-coding gene sequences for all except *M. thermoresistibile* were obtained from the EMBL database, Integr8. The *M. thermoresistibile* genome sequence was a series of contigs from an incomplete assembly provided by Christoph Grundner of Seattle BioMed. For all species except *M. thermoresistibile* (which had no annotated genome available), NTM proteins with >40% sequence similarity and >70% coverage versus the 179 Mtb targets were identified using a BLASTP search and screened using the negative filters above. To identify *M. thermoresistibile* targets, a BLASTN search was performed using the targets selected from the closely related species, *M. smegmatis*, against the 255 assembled contigs of *M. thermoresistibile*. Corresponding ORF boundaries were obtained by generating all possible forward- and reverse-complement ORFs using GETORF from EMBOSS from the *M. thermoresistibile* contigs and performing a BLASTP search of the Mtb targets against the calculated ORFs, retaining proteins with >40% sequence similarity and >70% coverage.

### 4.2 Gene-to-structure pipeline

PCR, cloning, sequencing, expression screening, scale-up, and purification of proteins were performed as described previously (37-39). Work was stopped on any target for which a structure of another protein with >70% sequence identity over >90% of the sequence was solved in this study, or a protein with >95% identity over >90% of the sequence was solved in a different study, except when the target had already yielded high-quality diffraction data.

### 4.3 Enzyme structure and active site comparisons

Structures were selected for active site comparisons by performing a cross-BLASTP search of all 614 mycobacterial protein sequences in the RCSB PDB (www.rcsb.org/pdb), and keeping all pairs with >25% sequence identity and >70% coverage (see Fig. 2). This search yielded 252 protein pairs, 214 of which were enzymes. From these mycobacterial enzyme pairs, 106 pairs were identified having known active site residues for the Mtb enzyme - either from a substrate-bound Mtb structure, or from a previously published sequence and structure alignment study (PDB and reference PubMed IDs are listed in Table S3) - and having a pair of PDB structures available in the same ligand binding state (*i.e.* same substrate or large co-factor bound, or no large ligand bound). There were no pairs in the PDB with between 42% and 55% overall sequence identity fulfilling these selection criteria, explaining the lack of data points in this range in Figures 3 and 4.

Overall structure comparisons were performed using Molsoft ICM (40). Two different Cα RMSD values were calculated: one by including >70% of the Mtb sequence in the alignment (shown in Fig. 3*A* and in Table S3), and a second by including as much of the Mtb sequence as possible (minimum 85%) in the alignment (Table S3). The first method excludes some disordered regions such as loops and termini, while the second method includes such regions to a greater extent, causing higher RMSD values.

The Mtb enzyme active site residues were defined as those within 4Å of substrate in a substrate-bound PDB structure. NTM homolog active site residues were defined as residues that align with Mtb enzyme active site residues using sequence and structure alignment in CHIMERA MATCHMAKER (41). In some cases, no Mtb substrate-bound structure was available, but active site residues had been previously identified using sequence and structure alignment with a structure of an ortholog. In all cases, the reference PDB structure used and associated PubMed reference number are listed in Table S3, along with active site residues selected for the enzymes. Only active sites involving a single protein chain were used.

Active sites were compared using three methods: (1) Cα RMSD, a measure of the average distance between common Cα backbone atoms, using Molsoft ICM, (2) side chain identity of aligned active site residues, and (3) optimized superpositions of pharmacophoric property distributions using Molsoft ICM siteSuperAPF, as previously described in detail (26). For the last method, the selected active site residues within each PDB structure were converted into continuous 3D atomic property fields representing seven pharmacophoric properties - hydrogen bond donor, hydrogen bond acceptor, lipophilicity, size, electronegativity, charge, aromaticity/hybridization - and the optimal superposition of these fields was found using a Monte Carlo minimization procedure, yielding a pseudo-energy ($E_{APF}$) for a given pair. Pseudo-energies are additive: larger pairs of identical active sites will have larger $|E_{APF}|$ values than smaller pairs. To convert $E_{APF}$ values to 0-100% similarity ($PS_{APF}$), $E_{APF}$ scores were normalized using the formula:

$$PS_{APF} = \frac{-\sqrt{E_{APF}(A,B)}}{\sqrt{E_{APF}(A,A) \times E_{APF}(B,B)}}$$

where $E_{APF}(A,A)$ and $E_{APF}(B,B)$ are self-comparisons of the active sites for Mtb (A) and non-Mtb homolog (B) enzymes. This is a different normalization method than previously described (26) and was used to better distinguish between closely related pairs rather than to detect weakly related pairs. The resulting $PS_{APF}$ score represents the fractional similarity of the active sites: if one site has half of the atoms missing but is otherwise identical to the other, $PS_{APF}$ would be 0.5, or 50%.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This research was funded under Federal Contracts HHSN272200700057C and HHSN272201200025C from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services. Research conducted at Pacific Northwest National Laboratory was performed in the Environmental and Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research program.

## References

[1]. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: Confronting the challenges of antibacterial drug discovery. Nature Reviews Drug Discovery. 2007; 6:29–40.

[2]. Liebeschuetz JW, et al. PRO_SELECT: Combining structure-based drug design and array-based chemistry for rapid lead discovery. 2 The development of a series of highly potent and selective factor Xa inhibitors. J Med Chem. 2002; 45:1221–32. [PubMed: 11881991]

[3]. Sundstrom, M.; Norin, M.; Edwards, A. Structural genomics and high throughput structural biology. Taylor & Francis; Boca Raton, FL: 2006.

[4]. Myler PJ, et al. The Seattle Structural Genomics Center for Infectious Disease (SSGCID). Infect Disord Drug Targets. 2009; 9:493–506. [PubMed: 19594426]

[5]. Moon AF, Mueller GA, Zhong X, Pedersen LC. A synergistic approach to protein crystallization: Combination of a fixed-arm carrier with surface entropy reduction. Protein Sci. 2010; 19:901–13. [PubMed: 20196072]

[6]. Derewenda ZS. It's all in the crystals…. Acta Crystallogr D Biol Crystallogr. 2011; 67:243–8. [PubMed: 21460442]

[7]. Raymond A, et al. Combined protein construct and synthetic gene engineering for heterologous protein expression and crystallization using Gene Composer. BMC Biotechnology. 2009; 9:37. [PubMed: 19383143]

[8]. Lorimer D, et al. Gene composer: Database software for protein construct design, codon engineering, and gene synthesis. BMC Biotechnology. 2009; 9:36. [PubMed: 19383142]

[9]. Jaroszewski L, et al. Genome pool strategy for structural coverage of protein families. Structure. 2008; 16:1659–67. [PubMed: 19000818]

[10]. Kuzniar A, van Ham R, Pongor S, Leunissen J. The quest for orthologs: Finding the corresponding gene across genomes. Trends in Genetics. 2008; 24:539–51. [PubMed: 18819722]

[11]. Chim N, et al. The TB Structural Genomics Consortium: A decade of progress. Tuberculosis (Edinb). 2011; 91:155–72. [PubMed: 21247804]

[12]. Ioerger TR, Sacchettini JC. Structural genomics approach to drug discovery for Mycobacterium tuberculosis. Curr Opin Microbiol. 2009; 12:318–25. [PubMed: 19481971]

[13]. Anderson WF. Structural genomics and drug discovery for infectious diseases. Infect Disord Drug Targets. 2009; 9:507–17. [PubMed: 19860716]

[14]. Savchenko A, et al. Strategies for structural proteomics of prokaryotes: Quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. Proteins. 2003; 50:392–9. [PubMed: 12557182]

[15]. Baugh LB, et al. Combining functional and structural genomics to sample the essential Burkholderia structome. PLoS One. 2013; 8:e53851. [PubMed: 23382856]

[16]. Webber SE, et al. Design of thymidylate synthase inhibitors using protein crystal structures: The synthesis and biological evaluation of a novel class of 5-substituted quinazolinones. J Med Chem. 1993; 36:733–46. [PubMed: 8459400]

[17]. Appelt K, et al. Design of enzyme inhibitors using iterative protein crystallographic analysis. J Med Chem. 1991; 34:1925–34. [PubMed: 2066965]

[18]. Cushman DW, Cheung HS, Sabo EF, Ondetti MA. Design of potent competitive inhibitors of angiotensin-converting enzyme. Carboxyalkanoyl and mercaptoalkanoyl amino acids. Biochemistry. 1977; 16:5484–91. [PubMed: 200262]

[19]. Ojo KK, et al. Transmission of malaria to mosquitoes blocked by bumped kinase inhibitors. J Clin Invest. 2012; 122:2301–5. [PubMed: 22565309]

[20]. Andries K, et al. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science. 2005; 307:223–7. [PubMed: 15591164]

[21]. Rost B. Twilight zone of protein sequence alignments. Protein Engineering. 1999; 12:85–94. [PubMed: 10195279]

[22]. Berman HM, et al. The Protein Data Bank. Nucleic Acids Research. 2000; 28:235–42. [PubMed: 10592235]

[23]. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986; 5:823–6. [PubMed: 3709526]

[24]. Sánchez R, Sali A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci USA. 1998; 95:13597–602. [PubMed: 9811845]

[25]. Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One. 2007; 2:e383. [PubMed: 17440619]

[26]. Totrov M. Ligand binding site superposition and comparison based on Atomic Property Fields: Identification of distant homologs, convergent evolution and PDB-wide clustering of binding sites. BMC Bioinformatics. 2011; 12(Suppl 1):S35. [PubMed: 21342566]

[27]. Skolnick J, Gao M. Interplay of physics and evolution in the likely origin of protein biochemical function. Proc Natl Acad Sci USA. 2013; 110:9344–9. [PubMed: 23690621]

[28]. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003; 333:863–82. [PubMed: 14568541]

[29]. Johnson KA. Role of induced fit in enzyme specificity: A molecular forward/reverse switch. J Biol Chem. 2008; 283:26297–301. [PubMed: 18544537]

[30]. Ducati RG, Breda A, Basso LA, Santos DS. Purine Salvage Pathway in Mycobacterium tuberculosis. Current Med Chem. 2011; 18:1258–75.

[31]. Villela AD, Sanchez-Quitian ZA, Ducati RG, Santos DS, Basso LA. Pyrimidine salvage pathway in Mycobacterium tuberculosis. Current Med Chem. 2011; 18:1286–98.

[32]. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 2003; 48:77–84. [PubMed: 12657046]

[33]. Griffin JE, et al. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog. 2011; 7:e1002251. [PubMed: 21980284]

[34]. Bertrand T, et al. Sugar specificity of bacterial CMP kinases as revealed by crystal structures and mutagenesis of Escherichia coli enzyme. J Mol Biol. 2002; 315:1099–110. [PubMed: 11827479]

[35]. Briozzo P, et al. Structures of Escherichia coli CMP kinase alone and in complex with CDP: A new fold of the nucleoside monophosphate binding domain and insights into cytosine nucleotide specificity. Structure. 1998; 6:1517–27. [PubMed: 9862805]

[36]. Ofiteru A, et al. Structural and functional consequences of single amino acid substitutions in the pyrimidine base binding pocket of Escherichia coli CMP kinase. FEBS J. 2007; 274:3363–73. [PubMed: 17542990]

[37]. Bryan CM, et al. High-throughput protein production and purification at the Seattle Structural Genomics Center for Infectious Disease. Acta Crystallogr F Struct Biol Cryst Commun. 2011; F67:1010–4.

[38]. Stacy R, et al. Structural genomics of infectious disease drug targets: The SSGCID. Acta Crystallogr F Struct Biol Cryst Commun. 2011; 67:979–84.

[39]. Choi R, et al. Immobilized metal-affinity chromatography protein-recovery screening is predictive of crystallographic structure success. Acta Crystallogr F Struct Biol Cryst Commun. 2011; F67:998–1005.

[40]. Abagyan RA, Totrov MM, Kuznetsov DA. ICM: A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. J Comp Chem. 1994; 15:488–506.

[41]. Meng EC, Pettersen EF, Couch GS, Huang CC, Ferrin TE. Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics. 2006; 7:339. [PubMed: 16836757]

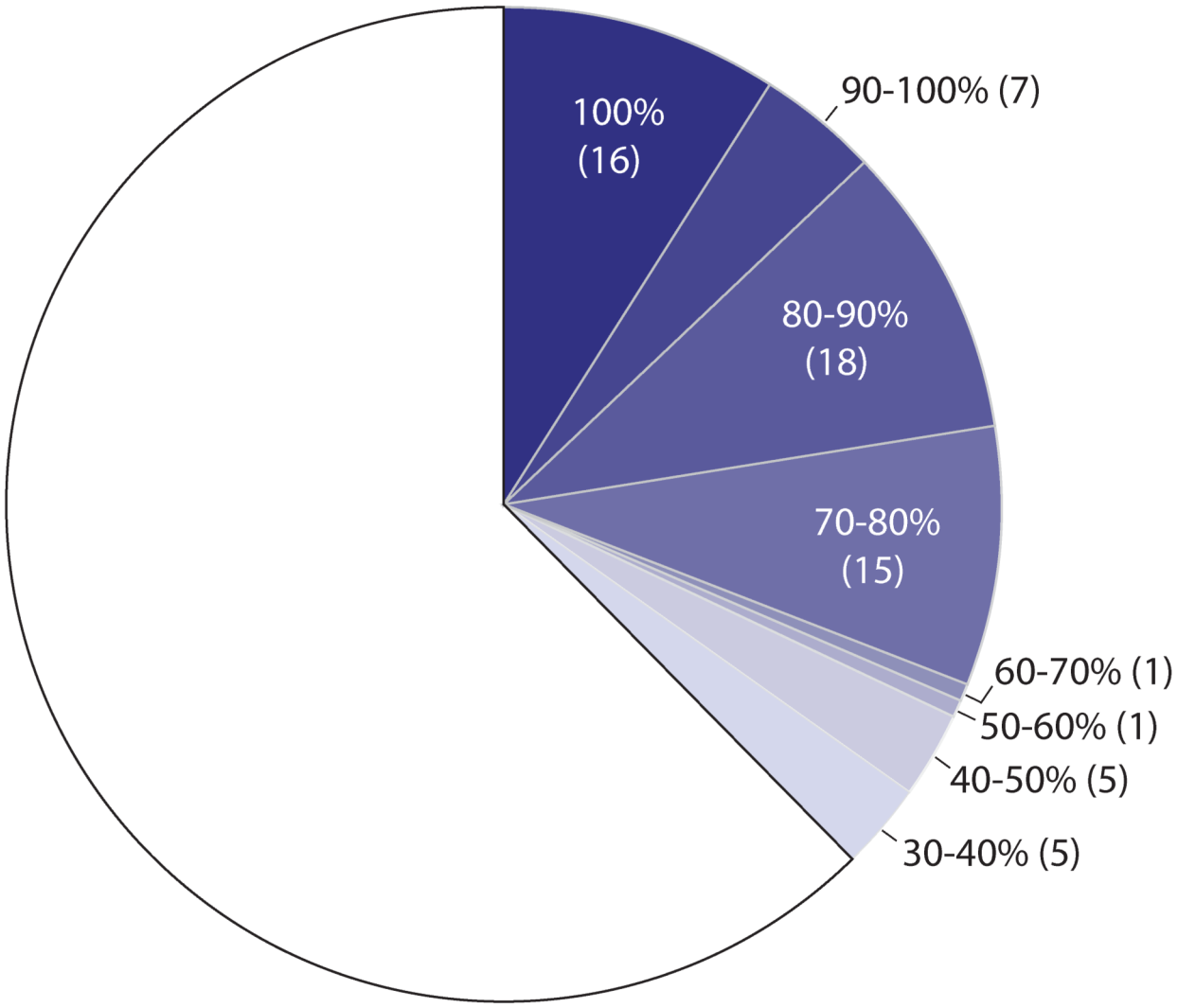


**Figure 1. Increasing the structural coverage of TB drug targets**

One hundred and seventy-nine Mtb proteins were selected based on their potential value as TB drug targets (represented by the entire circle). Structures were obtained for 16 of the Mtb proteins (100% amino acid sequence identity, dark blue). By adding 1675 mycobacterial homologs to the pipeline, structures of homologs of an additional 52 of the 179 Mtb targets were obtained (lighter shades of blue, corresponding to different levels of global sequence identity versus the Mtb target). In 42 cases where no structure was obtained for an Mtb protein, a structure of an NTM homolog with >55% sequence identity was solved.

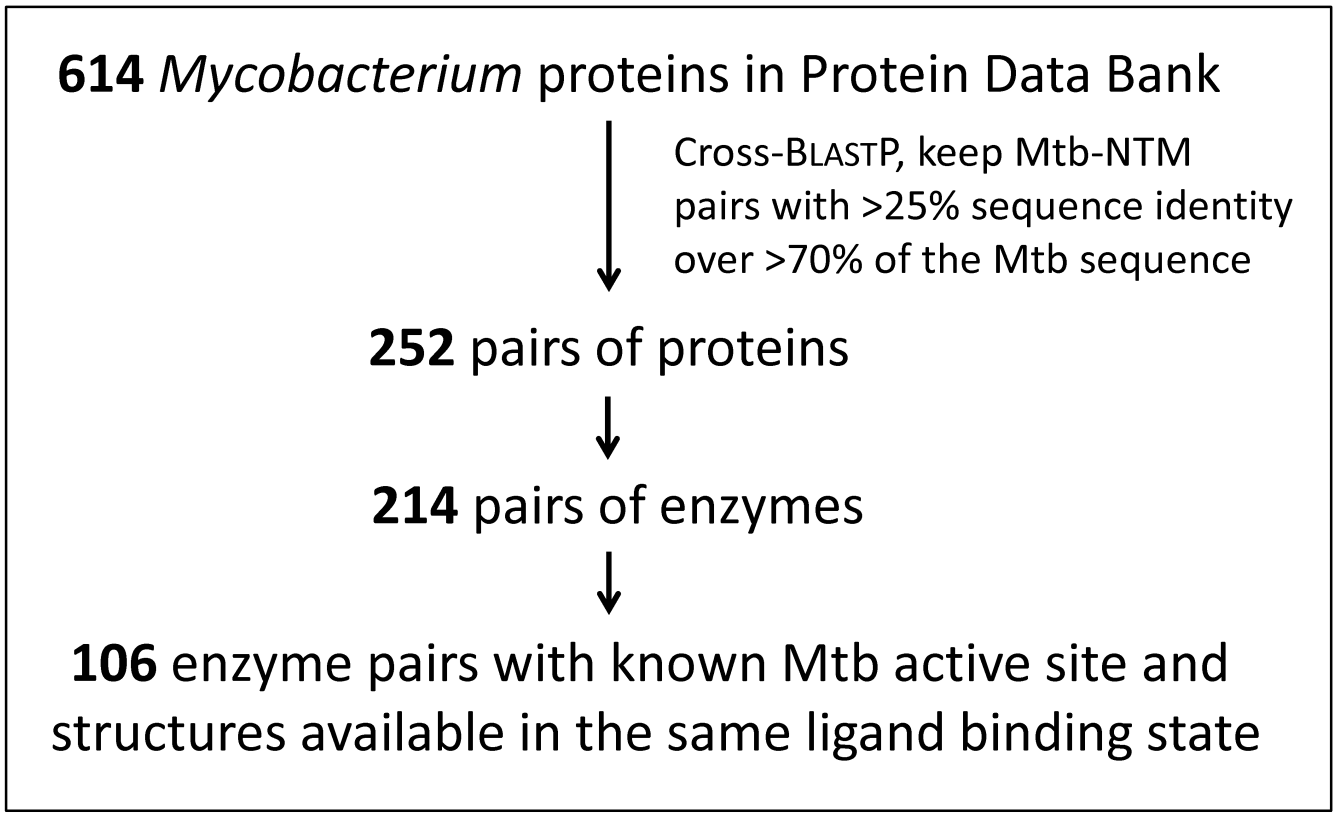


**Figure 2. Selection of mycobacterial enzyme homologs for active site comparisons**


**Figure 3. Comparison of enzyme overall and active site structure by Cα RMSD**

*A*, Overall RMSD between all Cα (backbone) atoms is plotted against global sequence identity for 106 Mtb and NTM enzyme structure pairs. The y-axis is inverted so that lowest RMSD values, which indicate greatest structural similarity, are at the top. *B*, Active site Cα RMSD is plotted against global sequence identity. Enzyme pairs in the same OrthoMCL family are indicated by filled data point markers, while pairs in different families are indicated by open markers.

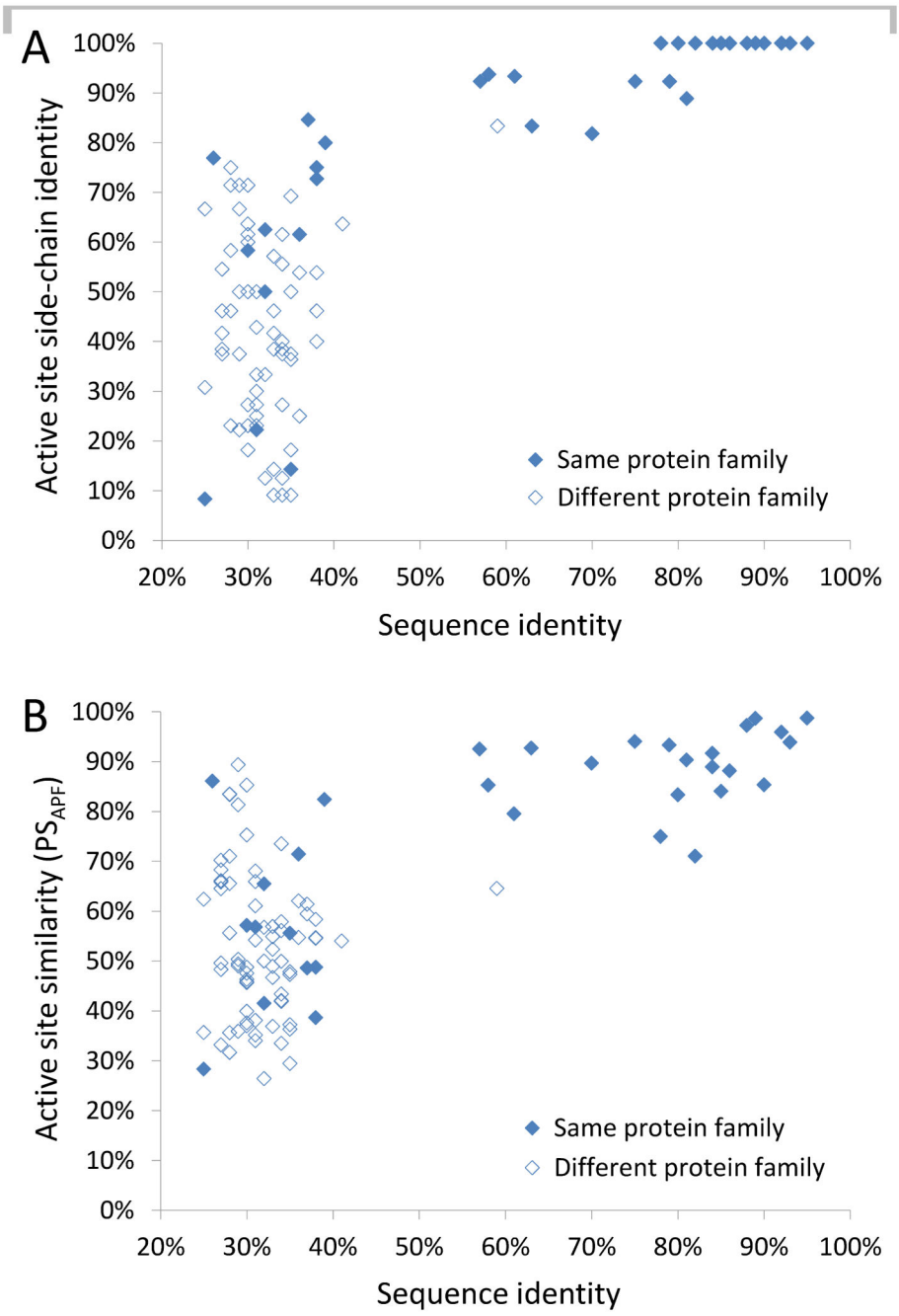


**Figure 4. Comparison of enzyme active site side chain identity and pharmacophoric properties**

*A*, Active site side chain identity is plotted versus global sequence identity for 106 Mtb and NTM enzyme pairs. *B*, Active site similarity based on optimized superpositions of pharmacophoric property distributions ($PS_{APF}$) is plotted against global sequence identity.

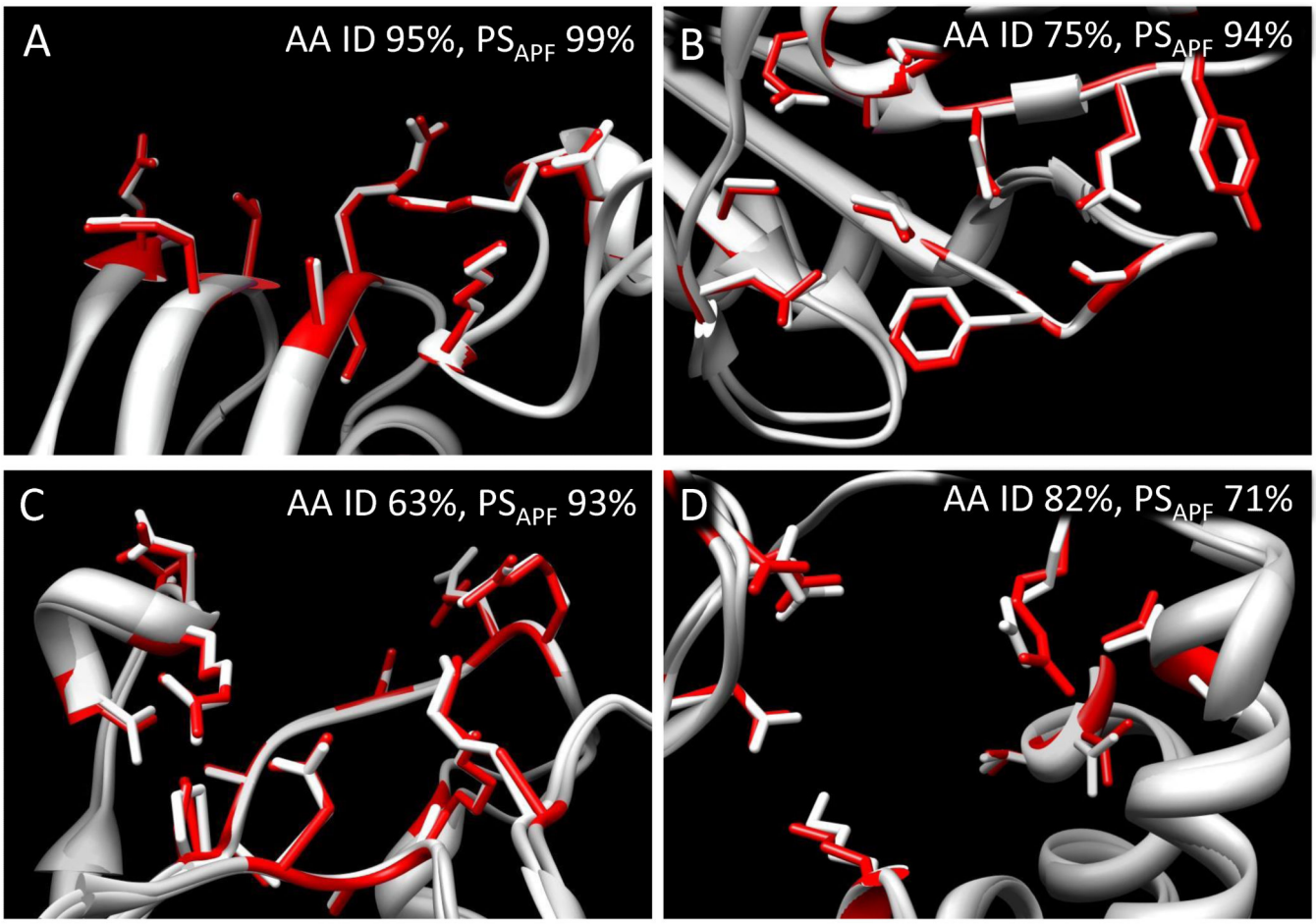


**Figure 5. Conserved active site structure for four mycobacterial enzymes**

*A*, metK from Mtb (3TDE) and *M. marinum* (3RV2). *B*, cdd from Mtb (3IJF) and *M. smegmatis* (3MPZ). *C*, ispD from Mtb (2XWN) and *M. smegmatis* (2XWL). *D*, gpgS from Mtb (3E25) and MAP_2569c from *M. paratuberculosis* (3CKQ). The Mtb structures are shown in white and gray, the NTM homolog structures in red. Global sequence identity (AA ID) and active site similarity score ($PS_{APF}$) are listed for each pair.

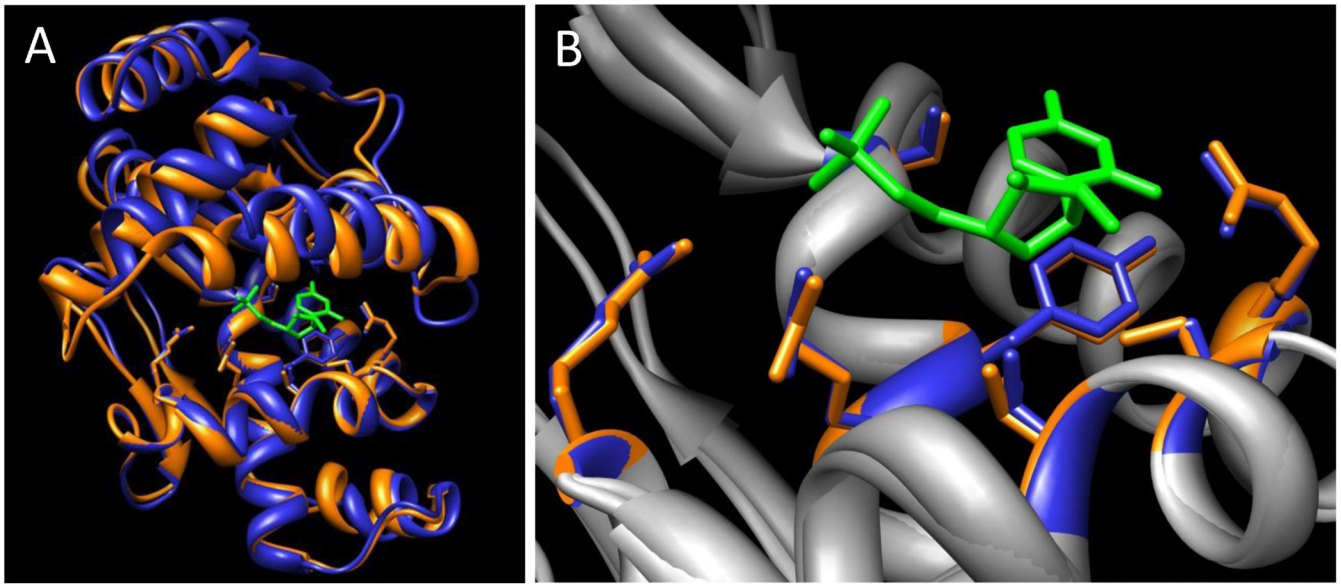


**Figure 6. Structural comparison between two homologs of cytidylate kinase, a potential TB drug target**

*A*, Homologs from *M. smegmatis* (3R20, orange) and *M. abscessus* (4DIE, blue), superimposed using active site chemical property distributions. The bound substrate, cytidine-5′-monophosphate (from 4DIE), is colored green. *B*, Enlarged view of the active sites with surrounding backbone structures in gray.

**Table 1**

Success rates by *Mycobacterium* species.

| Step | *tuberc-ulosis* | *abcessus* | *avium* | *bovis* | *leprae* | *marinum* | *paratub-erculosis* | *smegmatis* | *thermo-resistibile* | *ulcerans* | All | Rate by step |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Selected** | 141 | 188 | 231 | 22 | 70 | 200 | 185 | 239 | 74 | 151 | **1501** | |
| **Cloned** | 97% | 99% | 82% | 95% | 89% | 93% | 79% | 93% | 95% | 91% | **90%** | 90% |
| **Soluble** | 57% | 60% | 45% | 32%* | 27%** | 56% | 48% | 67%* | 53% | 32%** | **51%** | 57% |
| **Purified** | 45% | 51% | 33% | 23% | 16% | 39% | 37% | 53% | 36% | 21% | **39%** | 76% |
| **Crystallized** | 24% | 32% | 26% | 23% | 7% | 23% | 26% | 40%* | 27% | 12% | **26%** | 67% |
| **HQ data** | 13% | 13% | 11% | 0%* | 4% | 13% | 8% | 16% | 12% | 4% | **11%** | 42% |
| **In PDB** | 11% | 13% | 10% | 0% | 3%* | 12% | 7% | 13% | 11% | 3%** | **10%** | 88% |

Statistical comparison of the species-specific success rates was performed using Fisher's exact test and those with significant differences from Mtb are indicated by * ($P$<0.05) and ** ($P$<0.01).