A Homolog Rescue Approach Increases the Structural Coverage of Tuberculosis Drug Targets

Paul Wengender², Don Lorimer², Loren Baugh¹, Garry Buchko¹, Robin Stacy¹, Lance J. Stewarte, Thomas E. Edwards², Wesley C. Van Voorhis³, Peter J Myler⁴,5,6
Seattle Structural Genomics Center for Infectious Disease and ¹Seattle Biomedical Research Institute, Seattle, Washington; ²Beryllium, Bainbridge Island, Washington; Departments of ³Medicine, ⁴Global Health, ⁵Biomedical Informatics and Medical Education, ⁶Microbiology, Institute for Protein Design, University of Washington, Seattle, Washington; ⁷Pacific Northwest National Laboratory, Richland, Washington

INTRODUCTION

High-resolution three-dimensional structures of Mycobacterium tuberculosis (Mtb) proteins provide templates for tuberculosis drug design, but are available for only a small fraction of the Mtb proteome in part due to low x-ray crystallography success rates (∼10%). "Homologue rescue" – the use of homologues to obtain structures for difficult-to-crystallize targets – has been used previously, but information for homologue selection such as functional characterization or known orthologues is unavailable for many proteins. Here we establish a relationship between mycobacterial enzyme active site similarity and overall sequence identity by using large-scale enzyme active site comparisons. Active site comparisons for 106 pairs of mycobacterial enzymes showed that above 60% sequence identity, enzyme active site shape and chemistry are highly conserved. In crystal structure determination, the structural coverage of 179 potential tuberculosis drug targets was increased over three-fold, from 9% to 31%, with structures of homologues sharing >50% sequence identity from nine other Mycobacterium species. The results demonstrate the utility of a homologue rescue strategy for increasing the structural coverage of drug targets by using homologues selected from within the same genus.

Target Progress and Success Rates

| Table 1. Target progress and success rates. The table lists target counts, cumulative success rates (in parentheses), and overall success rates for each step (e.g. success rate of crystallization given that the target was purified) for all species (for right). |

<table>
<thead>
<tr>
<th>Target</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>55%</td>
</tr>
<tr>
<td>NTM</td>
<td>65%</td>
</tr>
<tr>
<td>Mtb</td>
<td>45%</td>
</tr>
</tbody>
</table>

Fig. 1. Increasing the structural coverage of Mtb targets. (A) Schematic of target selection and structure determination. One hundred seventy-nine potential Mtb drug targets and 1675 NTM homologs entered the structure determination pipeline, leading to structures representing 67 Mtb targets (86 Mtb targets – either a target for the Mtb protein itself (16 cases), or a structure for a homolog (51 cases)). (B) Distribution of structures representing the 67 Mtb targets in terms of sequence identity. If structures were solved for multiple homologs, only the structure with the greatest sequence identity to the Mtb target was counted. 100% means that a structure for the Mtb target itself was solved. (C) 2D network view of all targets selected and structural coverage of protein families. Mtb targets are arranged across the top with NTM targets below as shown in the insert, connected by short vertical lines representing BLASTP matches with >40% sequence similarity solved. Structures solved appear as long vertical lines. Four large protein superfamilies and 161 smaller families are represented by clusters of decreasing size, shifting vertically from left to right. An interactive version of this panel is available at www.ssgcid.org/publications/mtb-nmt-pdb-network.

Crystallography

Selection of Mycobacterium Enzymes for Structure Comparison

Fig. 2. Selection of Mycobacterium enzymes for structure comparisons. All Mtb and non-Mtb Mycobacterium enzymes having >25% sequence identity to known Mtb enzyme active site from a substrate-bound structure, and x-ray crystal structures available in the same ligand binding state were selected.

Comparison of Mycobacterium Enzyme Structures

Fig. 3. Structural comparison of Mycobacterium enzyme pairs by Co RMSD. The RMSD between Co backbone atoms in aligned Mtb and NTM enzyme structures is plotted against overall sequence identity. The y-axis is inverted so that lower RMSD values, which indicate greatest structural similarity, are at the top. (B) The RMSD between Co backbone atoms in the active site for each pair is plotted against sequence identity. Pairs of enzymes in the same OrthoMCL protein family are indicated by filled data point markers, while pairs in different protein families are indicated by open markers.

Comparison of active sites between Mycobacterium enzyme pairs. (A) Active site side-chain identity is plotted versus overall sequence identity for 106 enzyme pairs. (B) Active site pocket similarly based on optimized superposition of pharmacophoric property distributions (PS_PSSM) is plotted against overall sequence identity. PS_PSSM scores represent fractional atomic similarity. i.e. if one site has half of the atoms missing, it is otherwise identical to the other. The score would be 50%.

SUMMARY

1. When an X-ray crystal structure is not available for a Mtb target, high quality surrogate structures can often be obtained from homologues selected from other Mycobacteria species.
2. Sequence identity remains one of the few metrics available for homologous target selection. Above 60% sequence identity, enzyme function can be inferred from Enzyme Commission IDs. We expect that proteins with >60% sequence identity and identical EC numbers to have similar substrate-binding pockets.
3. Our comparison of Mycobacterium enzyme structures show that above 55% sequence identity, most pairs of enzymes possess nearly identical active sites based (1) backbone topology, (2) side-chain identity, and (3) chemical property.
4. Cytidylate kinase is one example for which no Mtb structure is available. Two homologues (73% ID, and 68% and 74% identity to the Mtb target) have side-chain orientations selected and structural coverage of drug targets by using homologues selected from within the same genus.

Fig. 4. Comparison of active sites between Mycobacterium enzyme pairs. (A) Active site side-chain identity is plotted versus overall sequence identity for 106 enzyme pairs. (B) Active site pocket similarity based on optimized superposition of pharmacophoric property distributions (PS_PSSM) is plotted against overall sequence identity. PS_PSSM scores represent fractional atomic similarity. i.e. if one site has half of the atoms missing, it is otherwise identical to the other. The score would be 50%.

ACKNOWLEDGEMENTS

The authors want to thank the following contributors without whom the work could not get done: Isabelle Phan, Darren Begley Matt Clifton, Brianna Armour, David Dranow, Brandy Taylor, Marvin Muruthi, Ian Abendroth, James Fairman, David Fox III, Shellei Dietricher, Bart Stabler, Anna Garberg, Ryan Cho, Stephen Hewitt, Alberto Napuli, Janette Myers, Lynn Barrett, Yang Zhang, Mician Ferel, Elizabeth Mundt, Katie Thompkins, Ngoc Tran, Sally Lyons-Abbott, Ariel Abramov, Aarthi Sekar and Dmitri Serbzhinskiy.

This project has been funded under Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract Nos. HHSN272201200025C and HHSN272200700057C.

Beryllium, Bainbridge Island, Washington; Departments of Medicine, Global Health, Biomedical Informatics and Medical Education, Microbiology, Institute for Protein Design, University of Washington, Seattle, Washington; Pacific Northwest National Laboratory, Richland, Washington

Fig. 5. Active site superpositions using atomic property fields. Mtb side-chains are shown in white, NTM side-chains in red, with surrounding backbone structure in gray. (A) metK from Mtb (3TDE) and M. marinum (3RV2), (B) cdd from Mtb (3UJF (37)) and M. smegmatis (3MPZ). (C) ispd from Mtb (2XW1 (31)) and M. smegmatis (2WKL). (D) gpgS from Mtb (3E25 (32)) and MAP_2569c from M. paratuberculosis (3CQK (33)). Overall sequence identity and pocket similarity score (PS_PSSM) are indicated at the top of each panel. Although the superpositions are based on atomic property fields that would be more accurately represented by space-filling spheres, it is easier to visualize side-chain orientations using the stick representations as shown.

Fig. 6. Surrogate structures for Mtb cytidylate kinase, a potential TB drug target. (A) Active site superposition of homologs from M. smegmatis (3RV2, orange) and M. abscessus (4DIE, blue), with bound substrate (cytidine-5'-monophosphate) from 4DIE indicated in green, using atomic property fields. (B) Enlarged view of the active site, with side-chains and substrate highlighted using the same color scheme, and surrounding backbone structures in gray.