DATABASE
The Journal of Biological Databases and Curation

Database, 2015, 1–10
doi: 10.1093/database/bav060
Original article

Original article

SInCRe—structural interactome computational resource for *Mycobacterium tuberculosis*

Rahul Metri[1,2], Sridhar Hariharaputran[1,3], Gayatri Ramakrishnan[2,4], Praveen Anand[1], Upadhyayula S. Raghavender[3], Bernardo Ochoa-Montaño[5], Alicia P. Higueruelo[5], Ramanathan Sowdhamini[3], Nagasuma R. Chandra[1], Tom L. Blundell[5] and Narayanaswamy Srinivasan[4,*]

[1]Department of Biochemistry and [2]Indian Institute of Science Mathematics Initiative, Indian Institute of Science, Bangalore, India, [3]National Centre for Biological Sciences, TIFR, UAS-GKVK Campus, Bellary Road, Bangalore, India, [4]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India, and [5]Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, UK

*Corresponding author: Tel: +91-80-22932837, Fax: +91-80-23600535 Email: ns@mbu.iisc.ernet.in

Citation details: Metri,R., Hariharaputran,S., Ramakrishnan,G., *et al*. SInCRe—structural interactome computational resource for *Mycobacterium tuberculosis*. *Database* (2015) Vol. 2015: article ID bav060; doi:10.1093/database/bav060

Received 23 January 2015; Revised 14 May 2015; Accepted 26 May 2015

Abstract

We have developed an integrated database for *Mycobacterium tuberculosis* H37Rv (Mtb) that collates information on protein sequences, domain assignments, functional annotation and 3D structural information along with protein–protein and protein–small molecule interactions. SInCRe (Structural Interactome Computational Resource) is developed out of CamBan (Cambridge and Bangalore) collaboration. The motivation for development of this database is to provide an integrated platform to allow easily access and interpretation of data and results obtained by all the groups in CamBan in the field of Mtb informatics. In-house algorithms and databases developed independently by various academic groups in CamBan are used to generate Mtb-specific datasets and are integrated in this database to provide a structural dimension to studies on tuberculosis. The SInCRe database readily provides information on identification of functional domains, genome-scale modelling of structures of Mtb proteins and characterization of the small-molecule binding sites within Mtb. The resource also provides structure-based function annotation, information on small-molecule binders including FDA (Food and Drug Administration)-approved drugs, protein–protein interactions (PPIs) and natural compounds that bind to pathogen proteins potentially and result in weakening or elimination of host–pathogen protein–protein interactions. Together they provide prerequisites for identification of off-target binding.

Database URL: http://proline.biochem.iisc.ernet.in/sincre

Introduction

Mycobacterium tuberculosis H37Rv (Mtb), a causative agent of tuberculosis (TB), has remained a major health concern globally. Based on the World Health Organization (WHO) latest reports, it is estimated that there have been 8.6 million new cases of TB reported in 2012 and a total of 1.3 million TB deaths (1). Most patients are treated for TB using first-line drugs, rifampicin and isoniazid. Together with the other first-line drugs, ethambutol and pyrazinamide, these two drugs form the basic ingredients of combination chemotherapy followed by the WHO directly observed treatment short course strategy (2). Second-line drugs, such as fluoroquinolones, and injectables like kanamycin, capreomycin and amikacin, are relied upon when the first-line drugs fail to control the disease. However in recent times, many antibiotic-resistant strains of *Mtb* have been reported. Multi-drug resistant TB is an Mtb strain resistant to rifampicin and isoniazid. Furthermore, acquisition of resistance towards fluoroquinolone, along with at least one of the injectable drugs, causes extensively drug resistant TB (3). The emergence of resistant strains to the first- and second-line drugs currently used poses a mammoth challenge for control of TB and cure of the infected.

Off-target effects, which are often discovered at the later stages of drug discovery research, have led to failure of many new medicines. Thus, there is an urgent need to discover ways of identifying off-target sites for drugs at an early stage in research. Detailed structural knowledge of the interactions between molecules in the cell provides one way of approaching this problem. The objective would be to define the structural interactome, an inventory of the various interactions between macromolecules and both natural and synthetic small molecules. The structural interactome can augment molecular-level interaction networks and provide a rich source of information on interactions between biological molecules and natural or synthetic ligands. Information on interactions between host and pathogen proteins will be helpful in identifying targets among pathogen proteins.

Integrated databases defining the structural interactome, bringing together information on protein sequences and structures, binding site properties, small molecules and their interactions, provide a valuable resource. Existing integrated databases on TB, TB Database (4, 5) and Tuberculist (6), provide information on genome, proteome, expression as well as corresponding references in the scientific literature but provide no information on structural interactomics comprising of binding sites, small molecules, druggability analysis of targets and functional domain assignments. This work from the CamBan (Cambridge—Bangalore) collaboration, involving four independent research groups from Cambridge and Bangalore, brings together various resources developed by these research groups and elsewhere to provide an extended Mtb structural-interactome resource. Each group has contributed towards the data specific to TB using in-house algorithms and databases developed and established individually over the years. The algorithms used to generate the data are designed to address and enrich sequence and structural data along with various small-molecule interactions. The database also incorporates systems-based analysis and provides list of high-confidence targets.

Sensitive profile-based techniques such as hmmscan of HMMER3.0 (7), Reverse PSI-BLAST (Reverse Position-Specific Iterative Basic Local Alignment Search Tool) (8) and HHblits (9) were used to achieve enhanced domain annotation for the proteins. Structural annotation of *M. tuberculosis* proteome (10) and CHOPIN (11) database provided structural data for many proteins. PocketDepth (12), PocketMatch (13) and PocketAlign (14) algorithms are used for binding site prediction and comparison. Protein domain analysis of unannotated genes was pursued using a computationally intensive bioinformatics pipeline called PURE (Prediction of Unassigned Regions) (15). The dataset from CREDO (16), a protein–ligand interaction database for drug discovery, TIMBAL (17), a database of small molecules disrupting protein–protein interactions and TIBLE (http://mordred.bioc.cam.ac.uk/tible/), a database of small molecules against *Mtb* and ligand-based off-target predictions, are connected in structural interactome computational resource (SInCRe). High-confidence drug targets derived from targetTB (18) have been included in the database. Drug targets have also been identified by a sequence-based approach with the help of FDA-approved drugs and are incorporated into the database. An interface has also been provided to integrate the data available from other external resources like STRING (19), STITCH (20) and Tuberculist (6). Future works from these groups will be mapped on to Mtb-specific dataset, and SInCRe will be updated on a regular basis.

Database

The integrated suite of databases was developed to provide detailed sequence and structure-based dimensionality to aid in drug-discovery pipelines. The data integrated here were obtained from the databases and webservers developed individually by the four research groups in CamBan. In-house algorithms and databases are the primary resources for the database. Figure 1 details the various data types.

Sequence-based analysis

Algorithms and datasets used.

The development of similarity-search procedures with the use of profiles such as Position Specific Scoring Matrices

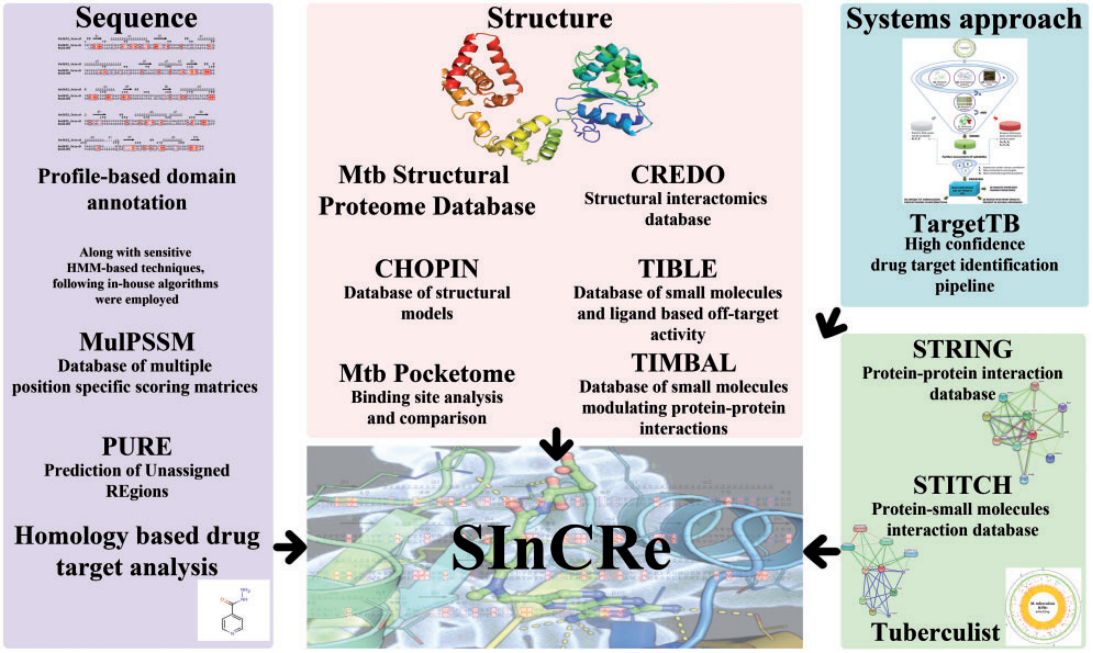

Figure 1. Various data resources contributing to SInCRe.

(PSSMs) (8, 21), Environment-Specific Substitution Tables (22) and Hidden Markov Models (HMMs) (7) has proven to be sensitive in detecting remote homologues reliably. Combination of such sensitive profile-based techniques resulted in the structural and functional annotations for ~95% of the Mtb proteome. Sensitive approaches such as hmmscan available through HMMER3.0 package, RPS-BLAST and HHblits were employed against sequence and structural profiles of various domain families obtained from Pfam (23), SUPERFAMILY (24), MulPSSM (25, 26) and the HH-suite database (ftp://toolkit.genzentrum.lmu.de/pub/HHsuite/databases/hhsuite_dbs/).

MulPSSM, developed by one of our research groups, is a searchable database of multiple PSSM profiles. The multiple profiles for a given domain family correspond to an alignment, wherein multiple sequences from that family are used as reference. The current version comprises 403 107 profiles for 14 831 Pfam domain families (Pfam v.27) and 14 235 profiles corresponding to 3856 structural families in PALI (Phylogeny and ALIgnment of homologous protein structures) (27) database based on SCOP (Structural Classification Of Proteins) (version 1.75).

Assessment of structure and function predictions.

Each of the associations made was assessed based on e-value and alignment length. For associations made using RPS-BLAST against MulPSSM database, an e-value cut-off of 0.01 was used in addition to a profile-coverage threshold of 70%. For domain assignments made using hmmscan against the Pfam database, profile-specific gathering threshold cut-offs were used to extract reliable hits. For hits identified by HHblits, an e-value threshold of 0.001 was used to associate domain families. In the searches against SUPERFAMILY HMM database using hmmscan software, the hits with e-values better than 0.0001 were considered to be reliable and are included in the database. Domain assignments for all the proteins were manually curated to maximize residue (64%) and sequence (89%) coverage.

Structural/functional domains for un-annotated proteins as well as those with unassigned regions were determined with help of the computationally intensive pipeline, PURE (15), developed by one of our groups. Cases where all the earlier approaches were unsuccessful in recognition of structural or functional domains, a fold-recognition algorithm PHYRE2 (28) was employed. A confidence cut-off of 90% was considered to retrieve folds reliably. This exercise was essentially an attempt to assess the foldability of the protein in question. Also, transfer of function based on homology was pursued using HHblits against non-redundant sequence database at an e-value cut-off of 0.001 and query coverage threshold of 60%.

All the hits of the earlier approaches were coupled with manual intervention to ensure maximum residue and sequence coverage of Mtb proteome (29).

Drug target identification based on sequence information.

Repurposing drugs has been regarded as a promising strategy mainly due to the reduced cost and time involved. A target identification methodology which essentially

integrates homology and pharmacological information (G. Ramakrishnan, N. Chandra and N. Srinivasan, in preparation) facilitated recognition of 132 FDA-approved drugs which could be repurposed for 56 potential targets in *Mtb*. This methodology comprises three steps: exploration of evolutionary relationship between targets of known FDA-approved drugs and *Mtb* proteins, structural elucidation of binding sites of *Mtb* proteins homologous to known targets and evaluation of predicted binding sites with the help of protein–ligand docking. Evolutionary relationships were explored with the help of a sensitive profile-based iterative search tool, jackhmmer(30), at an e-value threshold of 0.0001. An initial filtering step to eliminate drugs known to act on human proteins ensured that the 'anti-targets' in host are not picked up. The reliably identified relationships picked were further probed for the conservation of ligand-binding site residues across known targets and the *Mtb* proteins homologous to these targets. Structural information was taken from Protein Data Bank (PDB) (31) for *Mtb* and high-confidence structural models obtained from ModBase (32) for proteins with no known structure were used to assess the binding pockets. A structural alignment algorithm, TM-align (33), was effective in identification of highly similar local structural matches (TM-score > 0.50) between targets and their corresponding homologues in *Mtb*. Finally, the shortlisted proteins in *Mtb* predicted to serve as potential targets were evaluated using FDA-approved drugs with the help of an efficient protein–ligand docking tool, Glide (34–36) (http://www.schrodinger.com/Glide). A total of 132 FDA-approved drugs (Supplementary Table S1) were thus identified, which could be repurposed for 56 potential target proteins in *Mtb*.

Structure-based analysis

Structural proteome of Mtb.

Structural annotation of the *M. tuberculosis* proteome was carried out by one of the groups (10). PDB holds a total of 324 crystal structures of Mtb proteins and comparative models were generated for 2737 proteins, thus giving structure availability for 70% of the Mtb proteome. Structural models were generated using Modpipe, a software suite along with ModBase (32), a database of models generated using comparative modelling. The structural models need to be of high confidence and reliability as they play a central role to all the further analysis carried out. To assess the reliability of the protein structural models, various structure verification methods including statistical scoring potential (37, 38), secondary structure compatibility (39) and stereochemical quality check (40) were used. In the case of multi-domain proteins 3D models of individual domains are presented. Only those binding sites that were detected within the domains are analysed.

The CHOPIN (11) database (http://structure.bioc.cam.ac.uk/chopin) assigns structural domains and generates homology models for 2911 sequences, corresponding to ~73% of the proteome. Conformational states, characteristic of different oligomeric states and ligand binding, reflect various functional states of the proteins. Additionally, CHOPIN includes structural analyses of mutations potentially associated with drug resistance. The model number, sequence coverage and zscore are displayed on the SInCRe result page with links provided to CHOPIN webpage (http://mordred.bioc.cam.ac.uk/chopin/about) that provides model details and an option to download the models.

Detection of binding sites.

Computational methods for binding site detection can be classified into three broad categories based on their approaches: (i) evolutionary methods based on structure–sequence alignment (ii) energy-based methods using chemical probes and (iii) geometric approaches that scan the 3D structure of the protein to detect pockets. Each of these methods has its own strengths and limitations with respect to different aspects such as accuracy in detection and prediction, computational time, complexity and features captured. All the three methods were used in this study to minimize the prediction error and increase the confidence. The methods used are, a grid-based geometric method, PocketDepth (12), evolutionary method, Ligsite (41) and energy-based method, SiteHound. PocketDepth is an in-house method that uses depth-based clustering algorithm for detecting putative binding sites in the given protein structures. The idea that depth is defined by the centrality of empty subspaces in a protein structure is used to identify the pockets from all the protein structures. The PocketDepth algorithm was later combined with LIGSITEcsc, which uses Connolly's surface (42) to identify surface–solvent–surface events that involves grooves and then detects binding sites in a given protein by mapping the degree of conservation of the residues in the selected surface. All the pockets detected by PocketDepth that are within 5 Å radius of the predicted LIGSITEcsc pockets were selected. SiteHound (43), an energy method that searches for interaction zones favourable for a methyl probe within the protein, was used on all the pockets identified as a filter to fetch out final set of consensus ligand binding sites.

Other than the binding sites identified by these methods, pockets were also selected based on the experimentally characterized binding site residues in each protein in the proteome or in their homologues. This was done by

fetching entries from the database using respective general feature format files obtained from UniProt database (44). Possible binding sites were identified by scanning each protein sequence in the proteome with known binding motifs from the Prosite (45) database to make sure they were not missed out by other methods in the workflow (46). The binding sites detected can be viewed using Jmol plugin and also co-ordinates of these binding pockets can be downloaded in pdb format.

Drug binding site database and comparison.

DrugBank (47) and DrugPort were used to prepare a combined list of drugs or drug-like compounds; these included approved and experimental drugs and nutraceuticals. XML data files were obtained from these two databases and later parsed to extract information on proteins complexed with any of these drugs present in PDB. The binding sites were then extracted from these complexes. Residues of all atoms that lie within 4.5Å of any atom in the drug molecule were extracted as part of the binding site. Ten thousand six hundred and fifty-eight (from Drugbank) (Supplementary Table S2) + 2516 (from Drugport) (Supplementary Table S3) drug-binding sites were obtained from PDB through this process. High-confidence targets from Mtb were scanned using these known drug-binding sites, and also drug-binding sites were scanned for similarities against different binding site clusters.

Structural interactome.

The structural interactomics database CREDO (16) provides details of pairwise atomic interactions of intermolecular and intramolecular contacts between ligands and macromolecule for the structures in PDB. The PDB codes in the database are linked to the results of CREDO. This database stores interaction between atoms as structural interaction fingerprints as implemented by Deng *et al.* (48). Thirteen different interaction types such as hydrogen bonds, halogen bonds, carbonyl interactions and more are currently implemented in CREDO. Polypeptide-residue mapping is done onto UniProt. This allows identification of modified, non-standard or mutated proteins in the PDB compared with sequence in UniProt. Further, small-molecule and protein interaction details are provided in the database. Physico-chemical properties are calculated for all the small molecules in PDB and these properties are important for evaluating its drug-likeness. Topological similarities of the small molecules based on 2D and 3D descriptors are also retrieved from the database. With these data, CREDO provides major structural interaction details to study small-molecule binding properties. The PDB structures used as templates for building models in SInCRe are linked to the CREDO database.

Structure binding molecules.

TIMBAL (17), a database of small molecules disrupting protein–protein interactions, provides us with a list of small molecules relevant to the proteins of *Mtb*. Previously constructed by manual curation, now TIMBAL is automated to identify a list of protein–protein interaction modulators. The PPI targets and their orthologs are identified by UniProt identifiers. Small molecules related to these proteins are searched using UniProt identifiers in ChEMBL database. The homologues of known protein–protein interactions to the proteins in Mtb are identified and corresponding small molecules are listed. Totally 21 Mtb proteins are homologous to proteins in TIMBAL database corresponding to 11 targets.

Ligand-based off-target prediction and small-molecule data.

There are two main approaches to predict off-target activity. The structure-based approach relies on the similarity of the targets binding pockets, whereas the ligand-based approach connects targets based on the similarity of their ligands. The two methodologies complement each other (49). TIBLE (http://mordred.bioc.cam.ac.uk/tible/) collects small-molecule data (Minimal Inhibitory Concentration (MIC) for mycobacterium and binding to isolated Mtb targets) from the ChEMBL database (50) and the CDD (51). There are 75 Mtb targets with small-molecule binding data. For each of these targets, three independent algorithms—SEA (52), PharmMapper (53) and PASS (54) are used to derive off-target ligand-based predictions. Link from TIBLE to PharmMapper offers pharmacophore-matching platform for potential target identification. The details of small molecules and ligand-based off-target are integrated into the SInCRe database and also linked to the TIBLE page for detailed information.

Systems-based target identification

Identification of high confidence drug targets is a primary factor for efficient drug treatment. TargetTB (18), a comprehensive *in silico* target identification pipeline, was developed by one of the groups. The pipeline is built by incorporating network-based analysis of the protein–protein interactions, a flux-balance analysis of the reactome, phenotype-essentiality data derived from experiments, targetability assessment based on sequence and structure analysis using in-house novel algorithms. Initially proteins that are important for the survival of *Mtb* were identified using

flux balance and network analyses. Subsequently comparative genomics with the host was carried out. Finally the viability of a protein to be a potential drug target was assessed using novel methods for structural analysis of binding sites. Further, expression-data analysis, providing correlation and non-similarity measures of target proteins to gut flora proteins and also to 'anti-target' proteins in the host, was analysed extensively. Four hundred and fifty-one high-confidence entries were identified by this analysis pipeline. These short-listed targets have been further analysed through phylogenetic profiling against 228 pathogen genomes to identify antibiotic targets of broad spectrum especially those specific to TB. Target proteins significant to mycobacterial persistence and drug resistance mechanisms have also been analysed and reported. The details of the targets identified through TargetTB pipeline has been integrated into this database.

Other resources

External data from STRING (19), a database of known and predicted protein–protein interactions, STITCH (20), a database of protein–small molecule interactions and Tuberculist (6) for primary details about each Mtb protein are integrated into the SInCRe database.

Coverage of the *M. tuberculosis* proteome in the database

Our analysis of the repertoire of *M. tuberculosis* proteins, using a multitude of sensitive techniques, has generated a resource of information including structural and functional domain assignments, potential drug-targets and small-molecule binders including FDA-approved drugs. Figure 2 summarizes the percentage coverage achieved for *M. tuberculosis* proteins and indicates that 3495 of 4018 proteins could be associated with at least one functional domain (Pfam domain) assignment while 3131 proteins could either be associated with structural domains (SCOP domains) or with proteins of known structure. In terms of domain assignment alone, a total of 3566 proteins (89%) could be associated with at least one structural or functional domain. Due to the combined use of sensitive profile-based techniques, the percentage of *M. tuberculosis* proteins associated with functional domains is 3% higher than the annotations available in databases such as Pfam; and the percentage coverage achieved in terms of structural domains is 8% higher than the structural annotations available in databases such as SUPERFAMILY.

Systematic means to identify potential drug targets in *M. tuberculosis* has resulted in recognition of 498 high-confidence targets, constituting 12% of the proteome. The SInCRe database also includes information on protein–protein interactions within *M. tuberculosis* as documented in resources such as STRING. Approximately 23 000 known or predicted protein–protein interactions in *M. tuberculosis* are mediated by 2386 (59%) proteins.

Our attempt to integrate information from diverse resources provides a unified platform to explore and investigate the usefulness of a predicted target or a small molecule in the context of drug development and drug discovery for TB.

Database and web interface

The SInCRe database is created by integrating resources from various other databases for 4018 Mtb proteins. This database has been developed on the Linux-Apache-

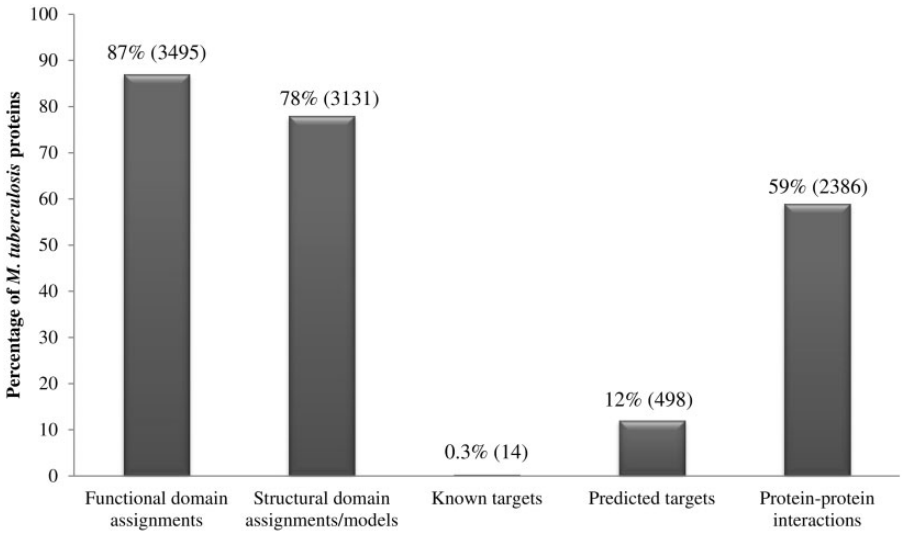

Figure 2. Percentage coverage of *M. tuberculosis* proteins in the database. Numbers in brackets denote absolute values.

MySQL-PHP platform. Sequence- and structure-level datasets have been stored in efficiently designed relational database schema. The web interface is developed using BootStrap (http://twitter.github.com/bootstrap). This provides cascading style sheets framework and javascript functionality. CytoscapeWeb (55), a java plugin, is used for interactive display of protein–protein and protein–small molecules interaction networks. Protein structures are represented in 3D using JSmol, a JavaScript-based molecular viewer from Jmol, an open-source Java viewer for chemical structures in 3D (http://www.jmol.org/). The modelled structures and sequences can be downloaded in PDB and FASTA formats, respectively. The tables in webpages are sortable and searchable, giving the user ease of acquiring data of interest.

The database can be queried using Rv IDs, gene name, UniProt ID, Pfam ID and Tuberculist functional classification. The dataset can be browsed for information available based on a few methods for limited list of Rv IDs.

Prediction of drug interactions using SInCRe

Protein kinases and phosphatases constitute important classes of drug targets due to the critical roles played by them in reversible protein phosphorylation that regulates many biological processes. There are many studies that report the development of potent inhibitors for these enzymes involved in protein phosphorylation to treat different types of cancer and autoimmune diseases (56). Serine/threonine protein kinases (STPKs) are one such class of kinases that specifically phosphorylate the hydroxyl group of one or more serine and threonine residues in the substrate protein. *Mycobacterium tuberculosis* (Mtb) genome houses 11 of such STPK genes and all of these are known to regulate crucial signalling processes, playing an important role in regulating physiology and virulence of the pathogen (57).

Of the 11 STPKs in Mtb, nine (PknA, PknB, PknD, PknE, PknF, PknH, PknI, PknJ and PknL) are receptors containing a transmembrane helix with extracellular sensory domain and intracellular kinase domain, thus acting as signal transducers. The other two kinases (PknG and PknK) are cytoplasmic containing a regulatory domain and could hence play a role in intracellular responses. Here, we explore the role of one such STPK – PknD (Rv0931c), as a putative drug target through the information present in SInCRe database. PknD acts a receptor kinase with extracellular sensory domain adopting a six-bladed β propeller structure (PDB ID: 1RWL, 1RWI) (58), and an intracellular kinase domain. The 3D structure of intracellular kinase domain could be derived using homology modelling using the crystal structure of PknE (PDB ID: 2H34) kinase domain as the template which share 59.7% sequence identity with the target. Although the substrate and the ligand for the PknD is yet to be discovered, the gene neighbourhood analysis reveals that it could play an essential role in phosphate transport. This is complemented by the fact that the growth of Δ*pknD* strain is compromised in a phosphate deficient medium (59). Recently, PknD has been observed to phosphorylate the N-terminal domain of Rv0516c, a putative regulator of sigma factor SigF (60). These three genes—PknD, Rv0516c and SigF play an important role in osmosensory signalling pathway (61). Moreover, a screen for identifying important genes for central nervous system infection by Mtb also identified PknD to be essential as Δ*pknD* strain was observed to be defective for invasion of central nervous system (62).

The binding site prediction exercise carried out on a proteome-scale involving a consensus of different types of algorithm (46) identified a putative binding site present at the interface of N-terminal and C-terminal lobe of kinase domain in PknD (Figure 3A). A systematic binding site comparison of this predicted pocket against a database of approved drug-binding sites yielded nilotinib (NIL) binding site from human mitogen activated protein kinase 11 protein (PDB ID: 3GP0) as the topmost hit with binding site similarity score (PMAX) (13) of 0.703. A binding site alignment of the predicted pocket with this known NIL binding site using PocketAlign algorithm (14) reveals the observed similarity and the differences in the binding sites (Figure 3B). Although the similarity of these protein kinases with the human counterparts can increase the risk of toxicity, there are supporting evidences in the literature that have successfully exploited the ATP-binding sites to achieve the selectivity. There are FDA-approved drugs that selectively bind to active and inactive conformations of the protein kinases to achieve the selectivity (56). The differences in kinase inhibitor binding sites (depicted as wireframe in Figure 3B) could be used as anchor points in fragment-based drug discovery to achieve the selectivity towards Mtb protein kinases. Interestingly, the binding sites of many of the anti-retroviral protease inhibitors like nelfinavir and lopinavir were also observed to have similarity to the predicted binding site in PknD. These observations are supported by the fact that nelfinavir is found to have anti-cancerous property attributed to its ability to weakly inhibit multiple protein kinases (63). One such anti-retroviral protease inhibitor—saquinavir (Ligand code: ROC), having high binding site similarity with the predicted binding site in PknD was explored further through computational docking using AutoDock Vina (Figure 3C) (64). The computationally predicted binding affinity (−8.1 kcal/mol) was found to be comparable to the native saquinavir complexed with HIV-protease (−9.4 kcal/mol). The best pose obtained through computational docking

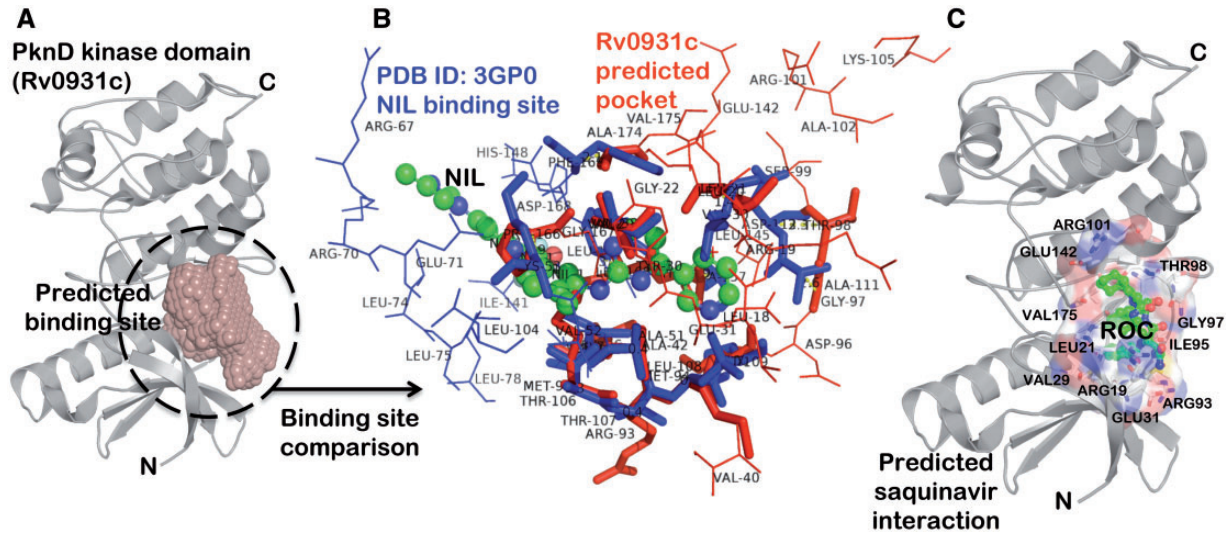

Figure 3. Example for predicted drug interactions using SInCRe. (**A**) A predicted binding site for PknD, a STPK, is depicted in the form of spacefill. (**B**) The alignment of predicted binding site from PknD (Rv0931c, in red) with the NIL binding site from Human Mitogen Activated Protein Kinase (PDB ID: 3GP0). The corresponding residues are highlighted in sticks, whereas unique residues with no correspondences are represented as wireframe. These distinguishing residues can be targeted to achieve the selectivity. (**C**) The best pose derived from computational docking depicting the interaction of saquinavir (ROC, shown as green ball and stick model) with the residues (represented as sticks) of the predicted binding site in PknD.

predicted the residues—ARG101, GLU142, ARG93 and GLU31 present in the predicted binding site to have crucial interaction with the saquinavir. These interesting drug associations can be readily obtained from the 'protein–small molecule associations' tab presented in the SInCRe database. The SInCRe database can thus, be used to generate readily testable hypothesis for anti-tubercular drug discovery.

Conclusion

SInCRe is an integrated suite of databases that provides the outcome of extensive sequence and structural studies of Mtb proteins. Sequence-based domain assignment and structural analysis of binding sites act as a resource to help in the identification off-target interactions of drug molecules, knowledge of which is useful in the design of novel drugs for *M. tuberculosis*. Future updates will include incorporation of other resources from Cambridge and Bangalore.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

Authors thank the two reviewers of this article, Samir Brahmachari, Anshu Bharadwaj and other colleagues from Council for Scientific and Industrial Research (CSIR), India for valuable comments and suggestions. They also thank Sumanta Mukherjee for his inputs on developing the database. Authors also thank Harry Jubb, John Overington, Yvonne Light, Sean Ekins, Xiaofeng Liu and John Irwin with their help in the development of the TIBLE resource.

Funding

This research is supported by Open Source Drug Discovery (OSDD) program of CSIR, India as well as by the Department of Biotechnology and Mathematical Biology Program sponsored. by Department of Science and Technology. N.S. is a J.C. Bose National Fellow. Funding for open access charge: National Centre for Biological Sciences (NCBS) to R. Sowdhamini.

Conflict of interest. None declared.

References

1. World Health Organization. (2013) Global Tuberculosis Report, **2013**. *World Health Organization, Geneva.*
2. Raviglione,M.C. and Uplekar,M.W. (2006) WHO's new stop TB strategy. *Lancet*, **367**, 952–955.
3. Dye,C. (2009) Doomsday postponed? Preventing and reversing epidemics of drug-resistant tuberculosis. *Nat Rev Microbiol*, **7**, 81–87.
4. Reddy,T.B., Riley,R., Wymore,F. *et al.* (2009) TB database: an integrated platform for tuberculosis research. *Nucleic Acids Res*, **37**, D499–D508.
5. Galagan,J.E., Sisk,P., Stolte,C. *et al.* (2010) TB database 2010: overview and update. *Tuberculosis*, **90**, 225–235.
6. Lew,J.M., Kapopoulou,A., Jones,L.M. *et al.* (2011) TubercuList–10 years after. *Tuberculosis*, **91**, 1–7.
7. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
8. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A. *et al.* (2002) CDD: a database of conserved domain alignments with

links to domain three-dimensional structure. *Nucleic Acids Res*, **30**, 281–283.
9. Remmert,M., Biegert,A., Hauser,A. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*, **9**, 173–175.
10. Anand,P., Sankaran,S., Mukherjee,S. *et al.* (2011) Structural annotation of *Mycobacterium tuberculosis* proteome. *PLoS One*, **6**, e27044.
11. Ochoa-Montaño,B., Mohan,N. and Blundell,T.L. (2015) CHOPIN: a web resource for the structural and functional proteome of *Mycobacterium tuberculosis*. *Database*, 2015, bav026.
12. Kalidas,Y. and Chandra,N. (2008) PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *J Struct Biol*, **161**, 31–42.
13. Yeturu,K. and Chandra,N. (2008) PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*, **9**, 543.
14. Yeturu,K. and Chandra,N. (2011) PocketAlign a novel algorithm for aligning binding sites in protein structures. *J Chem Inf Model*, **51**, 1725–1736.
15. Reddy,C.C., Shameer,K., Offmann,B.O. *et al.* (2008) PURE: a webserver for the prediction of domains in unassigned regions in proteins. *BMC Bioinformatics*, **9**, 281.
16. Schreyer,A.M. and Blundell,T.L. (2013) CREDO: a structural interactomics database for drug discovery. *Database*, **2013**, bat049.
17. Higueruelo,A.P., Jubb,H. and Blundell,T.L. (2013) TIMBAL v2: update of a database holding small molecules modulating protein-protein interactions. *Database*, **2013**, bat039.
18. Raman,K., Yeturu,K. and Chandra,N. (2008) targetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. *BMC Syst Biol*, **2**, 109.
19. Szklarczyk,D., Franceschini,A., Kuhn,M. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, **39**, D561–D568.
20. Kuhn,M., Szklarczyk,D., Pletscher-Frankild,S. *et al.* (2013) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res*, **42**, D401–D407.
21. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.
22. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, **310**, 243–257.
23. Punta,M., Coggill,P.C., Eberhardt,R.Y. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res*, **40**, D290–D301.
24. Gough,J., Karplus,K., Hughey,R. *et al.* (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol*, **313**, 903–919.
25. Anand,B., Gowri,V.S. and Srinivasan,N. (2005) Use of multiple profiles corresponding to a sequence alignment enables effective detection of remote homologues. *Bioinformatics*, **21**, 2821–2826.
26. Gowri,V.S., Krishnadev,O., Swamy,C.S. *et al.* (2006) MulPSSM: a database of multiple position-specific scoring matrices of protein domain families. *Nucleic Acids Res*, **34**, D243–D246.
27. Balaji,S., Sujatha,S., Kumar,S.S. *et al.* (2001) PALI-a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res*, **29**, 61–65.
28. Bennett-Lovsey,R.M., Herbert,A.D., Sternberg,M.J. *et al.* (2008) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins*, **70**, 611–625.
29. Ramakrishnan,G., Ochoa-Montano,B., Raghavender,U.S. *et al.* (2014) Enriching the annotation of *Mycobacterium tuberculosis* H37Rv proteome using remote homology detection approaches: insights into structure and function. *Tuberculosis*, **95**, 14–25.
30. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput Biol*, **7**, e1002195.
31. Berman,H.M., Westbrook,J., Feng,Z. *et al.* (2000) The protein data bank. *Nucleic Acids Res*, **28**, 235–242.
32. Pieper,U., Webb,B.M., Barkan,D.T. *et al.* (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*, **39**, D465–D474.
33. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, **33**, 2302–2309.
34. Halgren,T.A., Murphy,R.B., Friesner,R.A. *et al.* (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem*, **47**, 1750–1759.
35. Friesner,R.A., Banks,J.L., Murphy,R.B. *et al.* (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*, **47**, 1739–1749.
36. Friesner,R.A., Murphy,R.B., Repasky,M.P. *et al.* (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*, **49**, 6177–6196.
37. Colovos,C. and Yeates,T.O. (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci*, **2**, 1511–1519.
38. Shen,M.Y. and Sali,A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci*, **15**, 2507–2524.
39. Mereghetti,P., Ganadu,M.L., Papaleo,E. *et al.* (2008) Validation of protein models by a neural network approach. *BMC Bioinformatics*, **9**, 66.
40. Laskowski,R.A., Rullmannn,J.A., MacArthur,M.W. *et al.* (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, **8**, 477–486.
41. Huang,B. and Schroeder,M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol*, **6**, 19.
42. Connolly,M.L. (1993) The molecular surface package. *J Mol Graph*, **11**, 139–141.
43. Ghersi,D. and Sanchez,R. (2009) EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics*, **25**, 3185–3186.

44. Bairoch,A., Apweiler,R., Wu,C.H. *et al.* (2005) The universal protein resource (UniProt). *Nucleic Acids Res*, **33**, D154–D159.
45. Sigrist,C.J., de Castro,E., Cerutti,L. *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Res*, **41**, D344–D347.
46. Anand,P. and Chandra,N. (2014) Characterizing the pocketome of *Mycobacterium tuberculosis* and application in rationalizing polypharmacological target selection. *Sci Rep*, **4**, 6356.
47. Knox,C., Law,V., Jewison,T. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*, **39**, D1035–D1041.
48. Deng,Z., Chuaqui,C. and Singh,J. (2004) Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J Med Chem*, **47**, 337–344.
49. Xie,L. and Bourne,P.E. (2011) Structure-based systems biology for analyzing off-target binding. *Curr Opin Struct Biol*, **21**, 189–199.
50. Gaulton,A., Bellis,L.J., Bento,A.P. *et al.* (2011) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*, **40**, D1100–D1107.
51. Ekins,S., Kaneko,T., Lipinski,C.A. *et al.* (2010) Analysis and hit filtering of a very large library of compounds screened against *Mycobacterium tuberculosis*. *Mol Biosyst*, **6**, 2316–2324.
52. Keiser,M.J., Roth,B.L., Armbruster,B.N. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*, **25**, 197–206.
53. Liu,X., Ouyang,S., Yu,B. *et al.* (2010) PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res*, **38**, W609–W614.
54. Poroikov,V., Filimonov,D., Lagunin,A. *et al.* (2007) PASS: identification of probable targets and mechanisms of toxicity. *SAR QSAR Environ Res*, **18**, 101–110.
55. Lopes,C.T., Franz,M., Kazi,F. *et al.* (2010) Cytoscape web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
56. Zhang,J., Yang,P.L. and Gray,N.S. (2009) Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer*, **9**, 28–39.
57. Wehenkel,A., Bellinzoni,M., Grana,M. *et al.* (2008) Mycobacterial Ser/Thr protein kinases and phosphatases: physiological roles and therapeutic potential. *Biochim Biophys Acta*, **1784**, 193–202.
58. Good,M.C., Greenstein,A.E., Young,T.A. *et al.* (2004) Sensor domain of the *Mycobacterium tuberculosis* receptor Ser/Thr protein kinase, PknD, forms a highly symmetric beta propeller. *J Mol Biol*, **339**, 459–469.
59. Vanzembergh,F., Peirs,P., Lefevre,P. *et al.* (2010) Effect of PstS sub-units or PknD deficiency on the survival of *Mycobacterium tuberculosis*. *Tuberculosis*, **90**, 338–345.
60. Greenstein,A.E., MacGurn,J.A., Baer,C.E. *et al.* (2007) *M. tuberculosis* Ser/Thr protein kinase D phosphorylates an anti-anti-sigma factor homolog. *PLoS Pathog*, **3**, e49.
61. Hatzios,S.K., Baer,C.E., Rustad,T.R. *et al.* (2013) Osmosensory signaling in *Mycobacterium tuberculosis* mediated by a eukaryotic-like Ser/Thr protein kinase. *Proc Natl Acad Sci U S A*, **110**, E5069–E5077.
62. Be,N.A., Bishai,W.R. and Jain,S.K. (2012) Role of *Mycobacterium tuberculosis* pknD in the pathogenesis of central nervous system tuberculosis. *BMC Microbiol*, **12**, 7.
63. Xie,L., Evangelidis,T. and Bourne,P.E. (2011) Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput Biol*, **7**, e1002037.
64. Trott,O. and Olson,A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, **31**, 455–461.