NIH Public Access
Author Manuscript
Trends Microbiol. Author manuscript; available in PMC 2012 February 1.

Published in final edited form as:
Trends Microbiol. 2011 February ; 19(2): 65–74. doi:10.1016/j.tim.2010.10.005.

Computational databases, pathway and cheminformatics tools for tuberculosis drug discovery

Sean Ekins[1,2,3,4], **Joel S. Freundlich**[5], **Inhee Choi**[6,#], **Malabika Sarker**[7], and **Carolyn Talcott**[7]

[1]Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, PA 19046, USA

[2]Department of Pharmaceutical Sciences, University of Maryland, Baltimore, MD, USA

[3]Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine & Dentistry of New Jersey, Piscataway, New Jersey 08854, USA

[4]Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, CA 94010

[5]Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas 77843, USA

[6]Tuberculosis Research Section, Laboratory of Clinical Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

[7]SRI International, 333 Ravenswood Avenue, Menlo Park CA 94025, USA

Abstract

We are witnessing the growing menace of both increasing cases of drug-sensitive and drug-resistant *Mycobacterium tuberculosis* strains and the challenge to produce the first new tuberculosis (TB) drug in well over 40 years. The TB community, having invested in extensive high-throughput screening efforts, is faced with the question of how to optimally leverage this data in order to move from a hit to a lead to a clinical candidate and potentially a new drug. Complementing this approach, yet conducted on a much smaller scale, cheminformatic techniques have been leveraged and are herein reviewed. We suggest these computational approaches should be more optimally integrated in a workflow with experimental approaches to accelerate TB drug discovery.

New drugs for tuberculosis

Mycobacterium tuberculosis (*Mtb*), the causative agent of tuberculosis (TB), infects approximately one-third of the world's population and annually 1.7–1.8 million people die from this scourge [1]. The past decade has witnessed the growing menace of both increasing numbers of cases of drug-sensitive and drug-resistant strains and the recognition that fighting this global health pandemic, requires a multi-faceted research effort from both

© 2010 Elsevier Ltd. All rights reserved.

Corresponding author: Ekins, S. (ekinssean@yahoo.com).
[#]Current address: Medicinal Chemistry, Institut Pasteur Korea, Bundang-gu, Seongnamsi, Gyeonggi-do, Korea.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflicts of interest
S.E. is a consultant for Collaborative Drug Discovery. The other authors have no conflicts of interest.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

academia and industry. Infection with drug-sensitive TB may be handled with an existing front-line of four drugs. However, a lengthy treatment regimen (6–9 months typically), insufficient healthcare infrastructure especially in developing nations, and co-infection, with HIV/AIDS for example, often complicate the clinical scenario. Due to relatively low numbers of cases in the western hemisphere, TB is not a billion dollar blockbuster market that pharmaceutical companies are likely to profit from, and hence their involvement in research and development has been miniscule to date compared to cardiovascular disease, oncology, metabolic diseases, etc. This is especially important since the pipeline for TB therapeutics has not produced a new approved drug in over 40 years. Recently, phenotypic screening efforts have searched for compounds that inhibit the growth of *Mtb*, or its surrogate organisms such as *M. smegmatis* and *M. bovis* (BCG strain) [2]. These compounds broadly derive from chemically diverse libraries of small molecules, potentially providing the seeds for novel therapies. The TB community must now ask how to efficiently mine this growing database to afford new drug candidates, in the face of complications such as latency and persistence [3].

To help answer this question, we will turn to cheminformatics methods, which occupy an important place in the pharmaceutical industry drug discovery workflow. These computational approaches, in general, manage, mine and/or simulate complex systems or processes whether it is related to chemical, genomic, proteomic, or clinical data. Ligand- and protein-based methods, for example, have been used for the virtual screening of compound libraries as a complement to high-throughput screening *in vitro* [4]. Other researchers have described the different levels where computational approaches are used in drug discovery [5].

In the area of TB drug discovery, we review recently implemented computational approaches and resources that could be used to provide a roadmap for future efforts. Integration of these methods might guide the selection of compounds for additional *in vitro* screens and improve the odds of identifying new compounds as antitubercular hits or leads. While there have been several reviews on the current status of TB drugs and those in development [6] as well as isolated computational [7] and informatics-based [8] methods for drug discovery, to the best of our knowledge a review has not discussed the various computational tools [9] used in TB research. Others have suggested pipelines for bioinformatics processes such as target identification in TB (e.g. targetTB [10]) but there have been no suggested optimized and integrative cheminformatics workflows for antitubercular drug discovery.

Databases for TB

We are aware of over 300,000 compounds screened against *Mtb* in one laboratory alone, so it is likely that several million compounds have been examined cumulatively to date by all groups. It was not until recently that a central location for these screening results was developed. The advantage of collating data is that it might prevent repetition of screening by other groups while also allowing large-scale analysis of molecular properties of compounds with antitubercular whole-cell activity [11].

With so much data generated for different aspects of TB research, it is essential to have well curated databases. In Box 1, we summarize the range of some of the major databases for TB from diverse areas such as genome databases to databases of active compounds and refer the reader to the primary references and websites for further detail. Very few of the databases are linked so that a researcher can seamlessly navigate from one to another. We argue for greater levels of database connectivity or integration; a repository to point users to all these tools described in Box 1 is essential. These databases should be part of a workflow for TB

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

drug discovery such that the data are ultimately made available to the community once it is generated (Figure 1).

Pathway tools and technologies

It has been suggested that an integrated analysis of metabolic pathways, small molecule screening, and structural databases will facilitate anti-TB screening efforts [12], which reflects more of a systems biology (see Glossary) and computer-aided drug discovery approach. Predictive network-based systems biology approaches will be increasingly developed at the interface of cheminformatics and bioinformatics, with applications for target selection and discovery [13,14] alongside other target selection methods [15], areas of crucial importance to TB drug discovery. A translational systems biology approach to TB that integrates experimental and mathematical methods has also been proposed to bridge the isolated groups, and create collaborative groups of experimentalists and theoreticians [16].

Applications of systems biology to TB

One example of TB systems biology research is a study using gene expression data to identify stress response networks before and after treatment with different drugs [17]. The research combined the KEGG and BioCyc metabolic pathway databases with previously published gene expression data and a k-shortest path algorithm. It was found that gene expression networks for isoniazid treatment indicated a generic stress response. This type of approach could create an expression signature related to the drug used and the drug's mechanism of action [17].

A reaction influence network was created for *Mtb* using reactions as nodes, enabling protein-protein interaction mapping and identification of the putative consequences for global metabolism. For example, inhibition of Rv1653 (ArgJ) and Rv1131 (GltA1) could in turn maximally inhibit as much as 75% of metabolism [18]. A Boolean host-*Mtb* network model was also developed with 75 nodes representing molecules, cells, and processes that was used to simulate single and double *in silico* deletions [19]. KEGG and BioCYC pathway data (Box 2) were utilized as part of a domain fishing approach (using predominantly eukaryotic ligand-binding data) to generate compound-target networks as a means to deconvolute targets for 19 antitubercular agents without known target information [20].

A chemical systems biology approach can compare binding sites for known drugs and identified off-targets with similar binding sites. The FDA approved drugs entacapone and tolcapone, which target catechol-O-methyltransferase, were predicted to inhibit the enzyme enoyl-ACP reductase (InhA). Experimental data for entacapone showed that it has an MIC_{99} versus *Mtb* of 262 μM and inhibited InhA with an IC_{50} of 80 μM [21]. While these are very low potency hits perhaps quite distant from a starting point for drug discovery, they offer an intriguing path toward thinking about molecules that differ significantly from those previously known to target InhA.

Recently, the National Institute for Allergy and Infectious Diseases (NIAID) initiated a systems biology program (http://www.broadinstitute.org/annotation/tbsysbio/index.html) which aims to map the regulatory and metabolic networks of *Mtb* and the relevant state of these networks under conditions synonymous with TB pathogenesis. This will involve integration of profiling (multiple 'omics), high-throughput promoter mapping, bioinformatic and comparative sequence analysis, and computational modeling. While to date there have been relatively few applications of systems biology to TB, there is an opportunity to combine it with the field of cheminformatics, which has a far longer history with TB research.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Computational cheminformatic tools and their uses

Computational approaches applied to TB have predominantly implemented standard commercially available cheminformatic methods as will be described in the following section. These methods have been generally used by specialists focused on a single target or series of compounds and rarely in combination with other computational tools. Due to space limitations we have focused our analysis of cheminformatics tools used in TB research within the past 5 years as presented in the following sections and Tables 1–3.

Quantitative Structure-Activity Relationship and molecular properties analysis

Ligand-based approaches towards TB drug discovery primarily have used similar strategies over well over a decade. These approaches consist of Quantitative Structure-Activity Relationship (QSAR), 3D-QSAR, and pharmacophores. Once a model is generated using the appropriate, usually commercial software, testing is typically carried out by leaving out one or more groups of compounds at random. This is a very preliminary form of validation. Only rarely is an external test set generated after model building (see examples in Table 1).

These established ‘local’ models might help optimize antitubercular activity for a specific target or starting hit or lead (Table 1). In contrast, several analyses have utilized large datasets of active and inactive compounds tested against *Mtb* to calculate molecular descriptors or properties and analyzed for differences between the two groups (actives and inactives). Since many chemists and biologists are familiar with the ‘rule of 5’ [22] as a method for selecting ‘drug-like’ compounds, a significant question is whether anti-TB compounds obey Lipinski’s rules. This is often not the case. When 112 compounds known to have antitubercular activities [23] were filtered with the rule of 5, 40 (35.7%) compounds failed including known clinical candidates OPC-67683 and TMC-207 due to their lipophilicity and molecular weight. These clearly do not all represent approved drugs, and it remains to be seen if new TB drugs will fail this rule in the future. Analysis of several datasets representing many thousands of active compounds suggested that the mean value for various simple molecular descriptors, e.g. polar surface area (PSA), is statistically different from that of FDA approved drugs [11]. This analysis follows studies on molecular property values for antibiotics in general [25], including those that have evaluated logP and molecular mass [26], as well as earlier studies on antitubercular compounds [27]. Generally, FDA approved TB drugs are more like inhaled drugs (molecular weight mean 370, PSA 89.2 $Å^2$, clogP 1.7) [28]. An initial analysis of the largest public screening sets (over 300,000 compounds) to date using the MLSMR dataset [29] and the TAACF-NIAID-CB2 dataset [30] suggests the molecular weight, logP, and rule of 5 alerts were statistically significantly higher in the most active compounds of the MLSMR screening data, while the PSA is slightly lower compared to the inactive compounds. The active compounds in this TAACF-NIAID-CB2 set have statistically higher mean logP and rule of 5 alerts, while also having lower hydrogen bond donor count, atom count and PSA than inactive compounds [31]. These types of insights help define the ‘*Mtb*-active compound’ and can be used to design or select small molecule libraries for whole-cell phenotypic screens and to efficiently guide medicinal chemistry optimization efforts.

Comparative molecular field analysis and 3D-QSAR

As molecules interact with proteins in 3D, an understanding of molecular conformations for multiple molecules binding the same target provides useful information that can aid drug design. These methods could generate fields around the molecules and molecular descriptors based on conformation or a representation of a molecular feature that can then be related to bioactivity. These are termed 3D-QSAR. 3D-QSAR models (Table 2) have been generated with anywhere from 21 to ca. 100 molecules for narrow series of structurally related

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

compounds. In most cases, these studies have performed external testing on between < 10 to <30 compounds with generally good results. These models have rarely been used for anything other than data explanation with few virtual screening studies. There seem to be scant examples of global models generated using these methods, which likely stems from the limitations of comparative molecular field analysis (CoMFA) requiring rigid structural alignments [32], while other pharmacophore methods are generally alignment independent and can be used for rapid database searching [11]. Limitations of 3D-QSAR methods, in general, are the dependency on the molecule conformation, force fields, and the active compounds selected to build the model.

Classification machine learning methods

Machine learning and classification methods have been used sparingly for TB drug discovery. For example, the collation of 847 literature compounds and use of hologram QSAR allowed generation of fingerprint descriptors and clustering to identify features different between active and inactive compounds [33]. Such methods are valuable in the rapid virtual screening of compound libraries for novel actives. Models built with 60 or 71 molecules and up to 74 molecular descriptors were used to screen a library of 5000 compounds and discovered 18 active compounds [34]. Planche and co-workers used 122 compounds with fragment and topological substructural molecular design approach descriptors and linear discriminant analysis or *k*-means cluster analysis algorithms to predict the activity of a 2,4,5-trisubstituted imidazole class [35].

Classification methods can also be used as local models for lead optimization with smaller datasets. In one study 23 2,3-dideoxy hexenopyranosides were used with alignment free descriptors to generate combinatorial protocol multiple linear regression models that were tested by leaving out 8 compounds (r^2 0.64–0.74) [36].

The power of classification models has been demonstrated with several recent studies using much larger datasets. 3770 compounds collated by NIAID were used to build Bayesian classification models (cutoff MIC = 5 μM) with extended class fingerprints. The model was tested on a dataset of 2880 compounds (with activity against *Mtb*) from the GVKBio database with accuracy > 70% and was also used to screen the ZINC database, suggesting 4 compounds to be prioritized for future testing [37].

Bayesian models were built with the previously described MLSMR 220,463 library (4,096 active compounds) [30] and dose response data using 2,273 molecules (475 active compounds). In addition, these models implemented molecular function class fingerprints of maximum diameter 6 (FCFP_6) [38] and interpretable descriptors, and were tested [30] with the NIAID data and GVKBio datasets used by Prathipati *et al.* [37]. The models were further evaluated against the TAACF-NIAID-CB2 dataset of 102,634 molecules, resulting in a tenfold enrichment in compounds active against *Mtb* [31].These results indicate that classification methods could be used as computational filters prior to experimental testing. However, what is apparent from all the above studies is that prospective use and follow-up testing of suggested compounds is limited or non-existent to date.

Docking, virtual screening, and hybrid approaches

While many reports display images of *Mtb* protein binding sites to highlight interactions between ligand and protein, few have used them for computer-aided ligand design [39]. Docking is one such tool that can positively impact ligand or inhibitor design. Despite potential weaknesses due to under-sampling poses and how energetics are calculated through a scoring function, docking as a form of virtual screening has proven a useful tool outside the TB field [40].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Analyzing recent publications from 2007–2010 for this review indicates that docking has been utilized to identify small molecules with potency against a given *Mtb* target to: find hits, begin to build structure-activity relationships around early hits, and probe their metabolic stability. Docking has also been used as part of an integrated part of virtual screening processes and represents a complementary technology to biochemical high-throughput screening. Many use docking methods preceded by some form of computational filtering of screening libraries using pharmacophores or QSAR models (Table 3). Several of these studies have either suggested compounds for testing that have been validated or in several cases this validation is yet to be achieved. These hybrid methods can confirm the pharmacophore or QSAR model, with recent examples including thymidine analogs as inhibitors of thymidine monophosphate kinase (TMPK) [41]. Furthermore, in a search for InhA inhibitors, a 3D-QSAR-derived pharmacophore model was used to narrow down a set of 230,000 compounds to 299 top-scoring hits and ultimately 30 whose lowest energy docked conformation showed significant interactions with key active site residues, i.e. Tyr158, in addition to the 2'-hydroxyl of bound co-factor [42]. The predicted IC_{50} values were similar to the experimental values, although some of the molecules might be promiscuous binders.

Another study of TMPK inhibitors used a 3D-pharmacophore model derived from four X-ray structures of the enzyme with bound substrate or three inhibitors to screen a 60,000-compound vendor library [43]. Five of the eight virtual hits demonstrated whole-cell efficacy versus *Mtb*, but no TMPK inhibition data was presented. In a separate study, Nordqvist and colleagues searched for glutamine synthetase inhibitors using a combination of approaches [44]. A commercially available library of small molecules, chemically similar to substrate, product, or the known inhibitor *L*-methionine-*(S)*-sulfoximine, was virtually screened with scoring via a rigid pharmacophore model. After visual inspection, 4 of the 29 virtual hits demonstrated IC_{50} values of ~ 1 mM, which are very weak hits, but they facilitated design of a 15 member analog library as a starting point for future efforts.

All of these docking examples used a crystallographically characterized *Mtb* enzyme, yet others have utilized a homology model based on a closely related protein when a crystal structure is unavailable (e.g. work with UDP-N-acetylenolpyruvoylglucosamine reductase (MurB) [45] and fatty-acyl-CoA synthetase (FadD13) [46], which are involved in the biosynthesis of peptidoglycan and fatty acids, respectively). These efforts are dependent on the quality of the homology model and the extent of starting protein's similarity. Docking has also been utilized to investigate the metabolism of promising antitubercular small molecules. For example, the bioreduction of a nitro moiety in the BTZ043 family of inhibitors, which appears to target mycobacterial arabinogalactan and lipoarabinomannan polysaccharide biosynthesis, was studied [47]. Docking suggested potential BTZ043-*M. smegmatis* FMN-dependent nitroreductase NfnB interactions and proposed modifications to the BTZ043 scaffold to avoid metabolism via NfnB and other nitroreductases [47].

Docking, virtual screening, and hybrid approaches have resulted in some promising results, and yet, as discussed below, these methods and strategies require further significant refinements to eventually deliver on the promise of novel antitubercular therapeutics.

Gap analysis for computational methods in TB drug discovery

The computational methods previously described are widely used in the pharmaceutical industry in workflows by many project teams. We find several gaps when we look at how computational methods could be used in TB drug discovery (Figure 1) compared with the various reported efforts to date. Beginning with the recent popularity of high-throughput whole-cell phenotypic screening of large commercial libraries, we note limited use of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

filtering the library input or resulting hit lists for drug-likeness or lead-likeness [11]. Target deconvolution of the screening hits could clearly benefit from industry-derived computational methods [20]. When a follow-up screen is performed against a known biological target, virtual and biochemical screening could be performed sequentially. In seeking an eventual clinical candidate, we find only one mention of computational approaches for lead optimization to tackle issues with absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) [47]. This could be due to limited availability of global ADME/Tox models in academia compared with the pharmaceutical industry [32], and clearly represents an opportunity for impacting the quality of anti-TB compounds reaching the clinic in the future.

Successful use of computational approaches including virtual screening, docking and structure-based design are becoming more widespread in the pharmaceutical industry [48]. An example is Isentress™ (raltegravir), the first clinically approved HIV integrase inhibitor marketed by Merck, which was discovered using docking methods (AutoDock) and the relaxed complex method to accommodate receptor flexibility [49]. While we do not have this kind of a success story in TB drug development yet, it is hoped that computational technologies will have some visible impact and that this might be achieved by a greater realization of what is possible with readily available tools today.

Recommendations and outstanding questions

A disconnect in the TB community appears to exist between the generation and utilization of computational models and the entire drug discovery process. TB models are not well disseminated, shared or even reused and serve an isolated purpose for publication or comprehending a very limited structure-activity relationship. As it stands, these computational models reside in the hands of cheminformatics experts and insufficient efforts have been made in their dissemination on publicly accessible websites (in much the same way that databases are available and constantly accessible). For example, it could be feasible to use open technologies such as molecular descriptors and toolkits to generate TB or ADME/Tox models (perhaps derived from large pharmaceutical company datasets [50]) that could be shared with researchers regardless of affiliation. The linking of these tools to TB databases could begin to resolve this issue, analogous to the integration of technologies in systems biology.

With regard to integration, many examples are apparent in the TB literature that use combinations of computational approaches to improve potency. However, these still require integration within the drug discovery workflow (Figure 1) in which multiple iterations of many techniques are essential to move from hit to lead and beyond. It is widely accepted that enzyme inhibition (IC_{50} < 1 μM), whole-cell activity (MIC < 10 μM vs. *Mtb* H37Rv), and acceptable pharmacokinetic and toxicity profiles are necessary to facilitate study in animal models of infection, even before approaching clinical trials in humans.

The TB community is motivated to deliver novel therapeutics as quickly as possible. We suggest that computational workflows (Figure 1) could facilitate this and enable scientists to leverage these techniques at all stages of drug discovery as is common in the pharmaceutical industry. We hope this article promotes such an integrated use of computational techniques and collaborations across specialties within the TB field.

Box 1. TB related databases

BioHealthBase [51] is now incorporated into PATRIC (http://patricbrc.vbi.vt.edu/portal/portal/patric/IncumbentBRCs?page=bhb) and includes

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

rapid annotation using subsystem technology annotations for approximately 1,850 of the 2,000 complete bacterial genomes (including *Mtb*) currently available in PATRIC. The website provides a genome browser, protein family sorter, metabolic pathways (using KEGG pathway maps), phylogenetic trees, pathway and blast searches, feature cart, PubMed integration and Google search.

The Collaborative Drug Discovery Tuberculosis Database (CDD TB, www.collaborativedrug.com) [52] software (Collaborative Drug Discovery Inc. Burlingame, CA) is focused on small molecule libraries of compounds tested against *Mtb* [11]. CDD have collated over 15 public datasets on *Mtb* specific datasets representing well over 300,000 compounds derived from patents, literature and high throughput screening data shared by academic and pharmaceutical laboratories. In addition, this web-based database system [52] can facilitate storing and sharing of private data. The CDD database has been used to find compounds with molecular similarity to known *Mtb* drugs as well as build novel computational machine learning and pharmacophore models to rapidly identify potential inhibitors [11]. To date, CDD with funding from the Bill and Melinda Gates Foundation (BMGF) has developed a unique community with over 20 pilot groups in the TB field, including groups in the EU funded New Medicines 4 Tuberculosis (NM4TB) initiative [53] and groups funded by the BMGF Tuberculosis accelerator project.

GenoMycDB [54] is a database for the large-scale comparative analyses of completely sequenced mycobacterial genomes (http://157.86.176.108/~catanho/genomycdb/). It provides tools for functional classification and analysis of genome structure, organization, and evolution.

Tbrowse [55] is a resource for the integrative analysis of the TB genome, a genome browser across various online resources and datasets with over half a million data points (http://tbbrowse.osdd.net) and is a part of the open source drug discovery initiative (http://www.osdd.net/).

TDR targets database (http://tdrtargets.org) brings together genome sequencing and functional genomics projects, protein structural data, etc. [56]. Key features include computational assessment of target druggability and integration of large scale screening data with manually curated data, enabling the assembly of candidate targets to pursue.

Tuberculosis Drug Resistance Mutation Database [57] is a database listing mutations associated with TB drug resistance and the frequency of the most common mutations associated with resistance to specific drugs (http://www.tbdreamdb.com/).

TubercuList is widely recognized as the premier database for TB researchers. The TubercuList server [58] (http://genolist.pasteur.fr/TubercuList/help/about.html) represents a database focused on the analysis of the *Mtb* genomes as well as collating and integrating various aspects of the genomic information. TubercuList provides a complete dataset of DNA and protein sequences derived from *Mtb* H37Rv, linked to annotations and functional assignments.

The Tuberculosis Database (TBDB [59] http://www.tbdb.org/) provides genomic data (for 28 annotated genomes) and resources including several thousand microarray datasets from *in vitro* experiments and *Mtb* infected tissues. Researchers can freely deposit data prior to publication, browse gene detail pages, perform genome visualization and comparative analysis using the genome map tool, the genomes synteny map or operon map browser.

WebTB.org is provided by the TB structural genomics consortium [60–62]. It contains tools to search and browse the TB genome as well as structure summary pages on all

known TB proteins, the MTBreg database of proteins up- or downregulated in TB, top 100 persistence targets in TB and many more tools.

Box 2. Systems biology databases

BioCyc, MetaCyc (SRI) [63,64]: BioCyc (http://biocyc.org/MTBRV/) is a database collection together with a suite of tools supporting the generation of pathways and querying of them. The BioCyc database consists of organism-specific Pathway/Genome Databases (PGDBs), including tier 2 (derived computationally using the PathoLogic program as well as partially curated) PGDBs for two strains of *Mtb*, both virulent and drug-susceptible, namely CDC1551 and H37Rv. The PGDBs for *Mtb* are being adopted by the Tuberculosis Database (TBDB) [59] consortium (www.tbdb.org). This is expected to lead to more frequent updates reflecting the latest knowledge. The BioCyc collection also includes MetaCyc, a database of non-redundant, experimentally elucidated metabolic pathways curated from the experimental literature. MetaCyc (http://metacyc.org/) contains more than 1,200 pathways from more than 1,600 different organisms [65]. A PGDB describes the genome of an organism and the product of each gene; its metabolic network/pathways, reactions, enzymes, metabolites and transporter complement; and the genetic network of the organism, including its operons, transcription factors, and the interactions between transcription factors and their small-molecule ligands and DNA binding sites. The BioCyc Pathway Tools suite has three components. PathoLogic is used to create a new PGDB containing the predicted metabolic pathways of an organism, given an annotated genome, for example a Genbank entry, and MetaCyc as input. PathoLogic can predict metabolic pathways, genes coding for missing enzymes in metabolic pathways and operons. The Pathway/Genome Navigator supports query, visualization, and analysis of PGDBs. The Pathway/Genome Editors also allow interactive editing of PGDBs. In addition there is a computational interface to facilitate integration with external analysis tools such as the Pathway Tools Omics Viewer [66].

KEGG [67] is a major academic resource consisting of 16 databases covering genomic and chemical information and is a widely used reference resource (http://www.genome.jp/kegg/) valuable for linking compounds and metabolites to biological pathways [68,69].

LipidMaps [70] LIPID Metabolites And Pathways Strategy (LIPID MAPS) (http://www.lipidmaps.org/data/structure/LMSDSearch.php?Mode=SetupTextOntologySearch) was created in 2003 to identify and quantify all of the major and many minor lipid species in mammalian cells, as well as the changes in these species in response to perturbation.

Acknowledgments

S.E. acknowledges Dr. Barry A. Bunin and colleagues for developing the CDD TB database as well as the many TB research collaborators. The CDD TB database along with introductory training is provided freely to Mtb researchers through October 2010 thanks to funding from the Bill and Melinda Gates Foundation (Grant#49852 “Collaborative drug discovery for TB through a novel database of SAR data optimized to promote data archiving and sharing”). The project described was supported by Award Number R41AI088893 from the National Institute of Allergy and Infectious Diseases. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases or the National Institutes of Health. We sincerely apologize to any authors whose papers we omitted due to space restrictions.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

References

1. Balganesh TS, et al. Rising standards for tuberculosis drug development. Trends Pharmacol Sci 2008;29:576–581. [PubMed: 18799223]
2. Ballel L, et al. New small-molecule synthetic antimycobacterials. Antimicrob Agents Chemother 2005;49:2153–2163. [PubMed: 15917508]
3. Zhang Y. The magic bullets and tuberculosis drug targets. Annu Rev Pharmacol Toxicol 2005;45:529–564. [PubMed: 15822188]
4. Schneider G. Virtual screening: an endless staircase? Nat Rev Drug Discov 2010;9:273–276. [PubMed: 20357802]
5. Chandra N. Computational systems approach for drug target discovery. Expert Opin Drug Disc 2009;4:1221–1236.
6. Tomioka H. Current status of some antituberculosis drugs and the development of new antituberculous agents with special reference to their in vitro and in vivo antimicrobial activities. Curr Pharm Des 2006;12:4047–4070. [PubMed: 17100611]
7. Holton SJ, et al. Structure-based approaches to drug discovery against tuberculosis. Current protein & peptide science 2007;8:365–375. [PubMed: 17696869]
8. Scior T, Garces-Eisele SJ. Isoniazid is not a lead compound for its pyridyl ring derivatives, isonicotinoyl amides, hydrazides, and hydrazones: a critical review. Curr Med Chem 2006;13:2205–2219. [PubMed: 16918349]
9. Kantardjieff K, Rupp B. Structural bioinformatic approaches to the discovery of new antimycobacterial drugs. Curr Pharm Des 2004;10:3195–3211. [PubMed: 15544509]
10. Raman K, et al. targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC systems biology 2008;2:109. [PubMed: 19099550]
11. Ekins S, et al. A Collaborative Database And Computational Models For Tuberculosis Drug Discovery. Mol BioSystems 2010;6:840–851.
12. Zumla A, Grange J. Tuberculosis. BMJ (Clinical research ed 1998;316:1962–1964.
13. Ekins, S., et al. Systems biology: applications in drug discovery. In: Gad, S., editor. Drug discovery handbook. Wiley; 2005. p. 123-183.
14. Ekins, S.; Giroux, C. Computers and systems biology for Pharmaceutical Research and Development. In: Ekins, S., editor. Computer Applications in Pharmaceutical Research and Development. John Wiley and Sons; 2006. p. 139-165.
15. Hasan S, et al. Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. PLoS Comput Biol 2006;2:e61. [PubMed: 16789813]
16. Day J, et al. Tuberculosis research: going forward with a powerful "translational systems biology" approach. Tuberculosis (Edinburgh, Scotland) 2010;90:7–8.
17. Cabusora L, et al. Differential network expression during drug and stress response. Bioinformatics 2005;21:2898–2905. [PubMed: 15840709]
18. Raman K, et al. Strategies for efficient disruption of metabolism in *Mycobacterium tuberculosis* from network analysis. Molecular bioSystems 2009;5:1740–1751. [PubMed: 19593474]
19. Raman K, et al. A systems perspective of host-pathogen interactions: predicting disease outcome in tuberculosis. Molecular bioSystems 2010;6:516–530. [PubMed: 20174680]
20. Prathipati P, et al. Fishing the target of antitubercular compounds: in silico target deconvolution model development and validation. J Proteome Res 2009;8:2788–2798. [PubMed: 19301903]
21. Kinnings SL, et al. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol 2009;5:e1000423. [PubMed: 19578428]
22. Lipinski CA, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Del Rev 1997;23:3–25.
23. Chhabria M, et al. New frontiers in the therapy of tuberculosis: fighting with the global menace. Mini reviews in medicinal chemistry 2009;9:401–430. [PubMed: 19356120]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

24. Ghose AK, et al. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. J Phys Chem 1998;102:3762–3772.
25. O'Shea R, Moser HE. Physicochemical properties of antibacterial compounds: implications for drug discovery. J Med Chem 2008;51:2871–2878. [PubMed: 18260614]
26. Payne DA, et al. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat Rev Drug Disc 2007;6:29–40.
27. Barry CE 3rd, et al. Use of genomics and combinatorial chemistry in the development of new antimycobacterial drugs. Biochem Pharmacol 2000;59:221–231. [PubMed: 10609550]
28. Ritchie TJ, et al. Analysis of the calculated physicochemical properties of respiratory drugs: can we design for inhaled drugs yet? J Chem Inf Model 2009;49:1025–1032. [PubMed: 19275169]
29. Maddry JA, et al. Antituberculosis activity of the molecular libraries screening center network library. Tuberculosis (Edinburgh, Scotland) 2009;89:354–363.
30. Ananthan S, et al. High-throughput screening for inhibitors of *Mycobacterium tuberculosis* H37Rv. Tuberculosis (Edinburgh, Scotland) 2009;89:334–353.
31. Ekins S, et al. Analysis and hit filtering of a very large library of compounds screened against *Mycobacterium tuberculosis*. Mol Biosys 2010;6:2316–2324.
32. Kortagere S, Ekins S. Troubleshooting computational methods in drug discovery. J Pharmacol Toxicol Methods 2010;61:67–75. [PubMed: 20176118]
33. Prakash O, Ghosh I. Developing an antituberculosis compounds database and data mining in the search of a motif responsible for the activity of a diverse class of antituberculosis agents. J Chem Inf Model 2006;46:17–23. [PubMed: 16426035]
34. Garcia-Garcia A, et al. Search of chemical scaffolds for novel antituberculosis agents. J Biomol Screen 2005;10:206–214. [PubMed: 15809316]
35. Planche AS, et al. Design of novel antituberculosis compounds using graph-theoretical and substructural approaches. Mol Divers 2009;13:445–458. [PubMed: 19340599]
36. Saquib M, et al. C-3 alkyl/arylalkyl-2,3-dideoxy hex-2-enopyranosides as antitubercular agents: synthesis, biological evaluation, and QSAR study. J Med Chem 2007;50:2942–2950. [PubMed: 17542574]
37. Prathipati P, et al. Global Bayesian models for the prioritization of antitubercular agents. J Chem Inf Model 2008;48:2362–2370. [PubMed: 19053518]
38. Jones DR, et al. Computational approaches that predict metabolic intermediate complex formation with CYP3A4 (+b5). Drug Metab Dispos 2007;35:1466–1475. [PubMed: 17537872]
39. Willand N, et al. Synthetic EthR inhibitors boost antituberculous activity of ethionamide. Nat Med 2009;15:537–544. [PubMed: 19412174]
40. Kolb P, et al. Docking and chemoinformatic screens for new ligands and targets. Curr. Opin. Biotechnol 2009;20:429–436. [PubMed: 19733475]
41. Andrade CH, et al. 3D-Pharmacophore mapping of thymidine-based inhibitors of TMPK as potential antituberculosis agents. J. Comput. Aided Mol. Des 2010;24:157–172. [PubMed: 20217185]
42. Lu XY, et al. Discovery of potential new InhA direct inhibitors based on pharmacophore and 3D-QSAR analysis followed by in silico screening. Eur. J. Med. Chem 2009;44:3718–3730. [PubMed: 19428156]
43. Kumar A, et al. Knowledge based identification of potent antitubercular compounds using structure based virtual screening and structure interaction fingerprints. J Chem Inf Model 2009;49:35–42. [PubMed: 19063713]
44. Nordqvist A, et al. Evaluation of the amino acid binding site of *Mycobacterium tuberculosis* glutamine synthetase for drug discovery. Bioorg. Med. Chem 2008;16:5501–5513. [PubMed: 18462943]
45. Kumar V, et al. Identification of hotspot regions of MurB oxidoreductase enzyme using homology modeling, molecular dynamics and molecular docking techniques. J. Mol. Model. 2010 DOI: 10.1007/s00894-010-0788-3.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

46. Jatana N, et al. Molecular modeling studies of Fatty acyl-CoA synthetase (FadD13) from *Mycobacterium tuberculosis*-a potential target for the development of antitubercular drugs. J. Mol. Model. 2010 DOI: 10.1007/s00894-010-0727-3.
47. Manina G, et al. Biological and structural characterization of the *Mycobacterium smegmatis* nitroreductase NfnB, and its role in benzothiazinone resistance. Mol. Microbiol. 2010 doi 10.1111/j.1365-2958.2010.07277.
48. Kubinyi, H. Success stories of computer-aided design. In: Ekins, S., editor. Computer Applications in Pharmaceutical Research and Development. John Wiley and Sons; 2006. p. 377-424.
49. Schames JR, et al. Discovery of a novel binding trench in HIV integrase. J Med Chem 2004;47:1879–1881. [PubMed: 15055986]
50. Gupta RR, et al. Using open source computational tools for predicting human metabolic stability and additional absorption, distribution, metabolism, excretion, and toxicity properties. Drug Metab Dispos 2010;38:2083–2090. [PubMed: 20693417]
51. Squires B, et al. BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. Nucleic Acids Res 2008;36:D497–D503. [PubMed: 17965094]
52. Hohman M, et al. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. Drug Disc Today 2009;14:261–270.
53. Makarov V, et al. Benzothiazinones kill *Mycobacterium tuberculosis* by blocking arabinan synthesis. Science 2009;324:801–804. [PubMed: 19299584]
54. Catanho M, et al. GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. Genet Mol Res 2006;5:115–126. [PubMed: 16755503]
55. Bhardwaj A, et al. TBrowse: an integrative genomics map of *Mycobacterium tuberculosis*. Tuberculosis (Edinburgh, Scotland) 2009;89:386–387.
56. Aguero F, et al. Genomic-scale prioritization of drug targets: the TDR Targets database. Nat Rev Drug Discov 2008;7:900–907. [PubMed: 18927591]
57. Sandgren A, et al. Tuberculosis drug resistance mutation database. PLoS medicine 2009;6:e2. [PubMed: 19209951]
58. Cole ST. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. FEBS Lett 1999;452:7–10. [PubMed: 10376668]
59. Reddy TB, et al. TB database: an integrated platform for tuberculosis research. Nucleic Acids Res 2009;37:D499–D508. [PubMed: 18835847]
60. Terwilliger TC, et al. The TB structural genomics consortium: a resource for *Mycobacterium tuberculosis* biology. Tuberculosis (Edinburgh, Scotland) 2003;83:223–249.
61. Goulding CW, et al. The TB structural genomics consortium: providing a structural foundation for drug discovery. Curr Drug Targets Infect Disord 2002;2:121–141. [PubMed: 12462144]
62. Rupp B, et al. The TB structural genomics consortium crystallization facility: towards automation from protein to electron density. Acta crystallographica 2002;58:1514–1518.
63. Talcott C, et al. Pathway logic modeling of protein functional domains in signal transduction. Pacific Symposium on Biocomputing 2004:568–580. [PubMed: 14992534]
64. Caspi R, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 2006;34:D511–D516. [PubMed: 16381923]
65. Caspi R, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 2008;36:D623–D631. [PubMed: 17965431]
66. Paley S, Karp PD. The pathway tools cellular overview diagram and omics viewer. Nucleic Acids Res 2006;34:3771–3778. [PubMed: 16893960]
67. Anishetty S, et al. Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. Comput Biol Chem 2005;29:368–378. [PubMed: 16213791]
68. Beste DJ, et al. GSMN-TB: a web-based genome-scale network model of *Mycobacterium tuberculosis* metabolism. Genome Biol 2007;8:R89. [PubMed: 17521419]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

69. Marri PR, et al. Comparative genomics of metabolic pathways in *Mycobacterium* species: gene duplication, gene decay and lateral gene transfer. FEMS Microbiol Rev 2006;30:906–925. [PubMed: 17064286]
70. Sud M, et al. LMSD: LIPID MAPS structure database. Nucleic Acids Res 2007;35:D527–D532. [PubMed: 17098933]
71. Fernandes JP, et al. QSAR modeling of a set of pyrazinoate esters as antituberculosis prodrugs. Arch Pharm (Weinheim) 2010;343:91–97. [PubMed: 20099263]
72. Dolezal R, et al. N-benzylsalicylthioamides: highly active potential antituberculotics. Arch Pharm (Weinheim) 2009;342:113–119. [PubMed: 19137534]
73. Nayyar A, et al. Synthesis, anti-tuberculosis activity, and 3D-QSAR study of ring-substituted-2/4-quinolinecarbaldehyde derivatives. Bioorg Med Chem 2006;14:7302–7310. [PubMed: 16843663]
74. Macaev F, et al. Synthesis of novel 5-aryl-2-thio-1,3,4-oxadiazoles and the study of their structure-anti-mycobacterial activities. Bioorg Med Chem 2005;13:4842–4850. [PubMed: 15993090]
75. Ventura C, Martins F. Application of quantitative structure-activity relationships to the modeling of antitubercular compounds. 1. The hydrazide family. J Med Chem 2008;51:612–624. [PubMed: 18176999]
76. Andrade CH, et al. Fragment-based and classical quantitative structure-activity relationships for a series of hydrazides as antituberculosis agents. Mol Divers 2008;12:47–59. [PubMed: 18373208]
77. Sivakumar PM, et al. QSAR studies on chalcones and flavonoids as anti-tuberculosis agents using genetic function approximation (GFA) method. Chem Pharm Bull (Tokyo) 2007;55:44–49. [PubMed: 17202700]
78. Manvar AT, et al. Synthesis, in vitro antitubercular activity and 3D-QSAR study of 1,4-dihydropyridines. Mol Divers 2010;14:285–305. [PubMed: 19554465]
79. Shagufta, et al. CoMFA and CoMSIA 3D-QSAR analysis of diaryloxy-methano-phenanthrene derivatives as anti-tubercular agents. J Mol Model 2007;13:99–109. [PubMed: 16858589]
80. Aparna V, et al. 3D-QSAR studies on antitubercular thymidine monophosphate kinase inhibitors based on different alignment methods. Bioorg Med Chem Lett 2006;16:1014–1020. [PubMed: 16290929]
81. Hevener KE, et al. Quantitative structure-activity relationship studies on nitrofuranyl anti-tubercular agents. Bioorg Med Chem 2008;16:8042–8053. [PubMed: 18701298]
82. Nayyar A, et al. Synthesis, anti-tuberculosis activity, and 3D-QSAR study of 4-(adamantan-1-yl)-2-substituted quinolines. Bioorg Med Chem 2007;15:626–640. [PubMed: 17107805]
83. Nayyar A, et al. 3D-QSAR study of ring-substituted quinoline class of anti-tuberculosis agents. Bioorg Med Chem 2006;14:847–856. [PubMed: 16214351]
84. Kim P, et al. Structure-Activity Relationships of Antitubercular Nitroimidazoles. 2. Determinants of Aerobic Activity and Quantitative Structure-Activity Relationships. J Med Chem 2009;52:1329–1344. [PubMed: 19209893]
85. Biava M, et al. Antimycobacterial agents. Novel diarylpyrrole derivatives of BM212 endowed with high activity toward Mycobacterium tuberculosis and low cytotoxicity. J Med Chem 2006;49:4946–4952. [PubMed: 16884306]
86. Gupta RK, et al. Structure-based design of DevR inhibitor active against nonreplicating Mycobacterium tuberculosis. J Med Chem 2009;52:6324–6334. [PubMed: 19827833]
87. Kumar A, Siddiqi MI. CoMFA based de novo design of pyrrolidine carboxamides as inhibitors of enoyl acyl carrier protein reductase from Mycobacterium tuberculosis. J Mol Model 2008;14:923–935. [PubMed: 18626672]
88. Kumar A, Siddiqi MI. Receptor based 3D-QSAR to identify putative binders of *Mycobacterium tuberculosis* Enoyl acyl carrier protein reductase. J Mol Model 2010;16:877–893. [PubMed: 19779936]
89. Kumar A, et al. New molecular scaffolds for the design of Mycobacterium tuberculosis type II dehydroquinase inhibitors identified using ligand and receptor based virtual screening. J Mol Model 2010;16:693–712. [PubMed: 19816720]
90. Banfi E, et al. Antifungal and antimycobacterial activity of new imidazole and triazole derivatives. A combined experimental and computational approach. The Journal of antimicrobial chemotherapy 2006;58:76–84. [PubMed: 16709593]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

91. Andrade CH, et al. Rational design and 3D-pharmacophore mapping of 5'-thiourea-substituted alpha-thymidine analogues as mycobacterial TMPK inhibitors. J Chem Inf Model 2009;49:1070–1078. [PubMed: 19296716]
92. Labello NP, et al. Quantitative three dimensional structure linear interaction energy model of 5'-O-[N-(salicyl)sulfamoyl]adenosine and the aryl acid adenylating enzyme MbtA. J Med Chem 2008;51:7154–7160. [PubMed: 18959400]
93. Wahab HA, et al. Elucidating isoniazid resistance using molecular modeling. J Chem Inf Model 2009;49:97–107. [PubMed: 19067649]
94. Cho Y, et al. Discovery of novel nitrobenzothiazole inhibitors for Mycobacterium tuberculosis ATP phosphoribosyl transferase (HisG) through virtual screening. J Med Chem 2008;51:5984–5992. [PubMed: 18778048]
95. Kumar M, et al. In silico structure-based design of a novel class of potent and selective small peptide inhibitor of *Mycobacterium tuberculosis* dihydrofolate reductase, a potential target for anti-TB drug discovery. Mol Divers. 2009
96. Hegymegi-Barakonyi B, et al. Signalling inhibitors against *Mycobacterium tuberculosis*--early days of a new therapeutic concept in tuberculosis. Curr Med Chem 2008;15:2760–2770. [PubMed: 18991635]
97. Gopalakrishnan B, et al. A virtual screening approach for thymidine monophosphate kinase inhibitors as antitubercular agents based on docking and pharmacophore models. J Chem Inf Model 2005;45:1101–1108. [PubMed: 16045305]
98. Lin TW, et al. Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 2006;103:3072–3077. [PubMed: 16492739]
99. Metaferia BB, et al. Synthesis of natural product-inspired inhibitors of *Mycobacterium tuberculosis* mycothiol-associated enzymes: the first inhibitors of GlcNAc-Ins deacetylase. J Med Chem 2007;50:6326–6336. [PubMed: 18020307]
100. Srivastava SK, et al. NAD+dependent DNA Ligase (Rv3014c) from *Mycobacterium tuberculosis*. Crystal structure of the adenylation domain and identification of novel inhibitors. J Biol Chem 2005;280:30273–30281. [PubMed: 15901723]

Glossary

Classification models	this technique enables analysis of very large structurally diverse training sets which learn to discriminate between active and inactive compounds.
Comparative Molecular Field Analysis (CoMFA)	this modeling method uses 3D descriptors or fields and their position to describe antitubercular activity.
Docking	this is the computational determination of the most energetically feasible poses of a small molecule in a protein binding site. When used in virtual screening, a list of top-ranked compounds is queried with regard to putative interactions that could explain the rankings.
Global model	this usually describes a QSAR model composed of a structurally diverse training set and might represent larger, more general models useful for predicting across different structures. These models are generally better at extrapolation and cover a wider chemical space.
Lipophilicity	this is most typically quantified as an estimated logP such as AlogP [24] or clogP where logP is defined as the log ($P_{octanol}/P_{water}$) and P is a partition coefficient for a given compound in a specific solvent.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Local model	this generally describes a QSAR model composed of a structurally similar training set, representing a smaller model for lead optimization. These models generally cannot extrapolate outside of a single chemical series and cover a narrow chemical space.
Machine learning	this represents various computational methods that can understand patterns in large datasets and learn to enable decision-making. Examples include classification models (e.g. decision trees, support vector machines, Bayesian methods etc).
Pharmacophore	this is frequently a type of 3D-QSAR or arrangement of key molecular features important for biological activity (e.g. hydrogen bonding, hydrophobic, charged regions or fields). Some pharmacophores represent the key features without any quantitative calculation and can be used for virtual screening of 3D databases [11].
Quantitative structure-activity relationship (QSAR)	relates antitubercular activity to molecular descriptors using an algorithm.
3D-QSAR	is a model that relates antitubercular activity to molecular descriptors or fields.
Systems biology	this is an emerging, cross-disciplinary field that endeavors to comprehend how the molecular components of life function together to create complex biological systems. It is usually represented by computational integration of very large quantities of genomic, proteomic, and metabolomic information captured from underlying pre-existing databases (Box 2). A wide spectrum of approaches to systems modeling exists including: (i) statistical analysis of large datasets, (ii) models of system kinetics, (iii) flux balance techniques, (iv) evolutionary models of drug resistance, and (v) symbolic models of processes.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

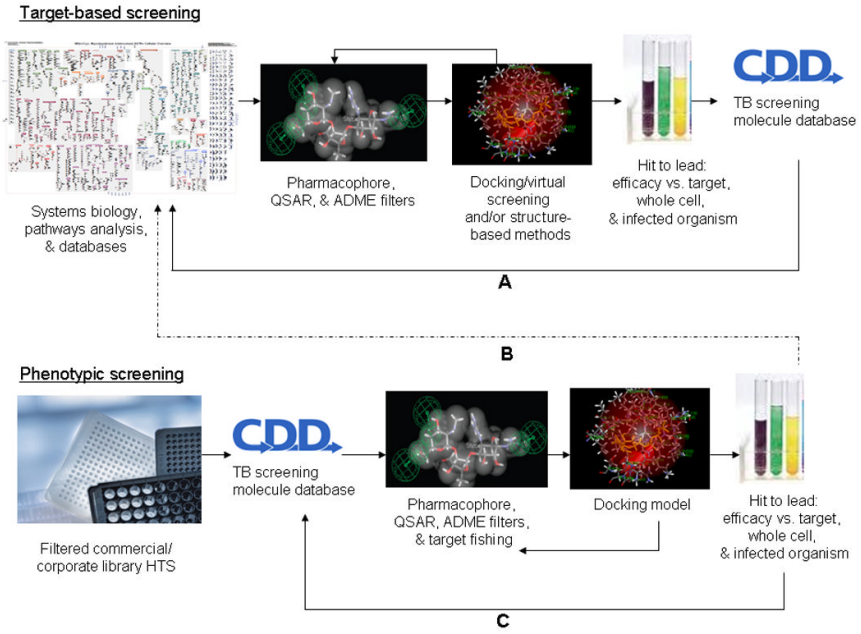

Figure 1. Workflows for target-based and phenotypic screening using multiple integrated computational components

Illustration of target-based screening to find new compounds that inhibit an enzyme or protein-protein interaction using tightly integrated computational methods, then optimize and feedback data to databases and pathways. Phenotypic screening data is used with integrated computational methods to suggest potential targets and optimize ADME properties in parallel, and then verify *in vitro*. Target-based screening computational methods may include: identification of target family interaction motifs; filtering and prioritization of compound source pools; design or selection for final screening collection; diversity, similarity and coverage calculation; and 2D or 3D descriptors (pharmacophore, shape, or chemical substructure). Target-based screening could use structure based methods and these could incorporate the following computational methods: 2D and 3D descriptor and pharmacophore based activity models; binding site assessment and mapping; ligand docking or virtual library screening; protein homology modeling; and fragment-based drug discovery. In both target based screening and phenotypic screening, hit to lead screening data analysis and follow-up might require computational tools for: 'hit-picking' and filtering, clustering, and prioritizing; isostere selection; identifying structure-activity relationship trends; and calculating chemical substructures and properties, e.g. 2D or 3D descriptors. Phenotypic screening might require computational methods for hit explosion (such as the creation of a pharmacophore or by similarity searching in commercial data bases) as well as target fishing [20] to identify the target for a hit. Lead optimization requires the use of computational methods for identifying, tracking, and optimizing structure-activity relationships and ADME trends within data sets. **(a)** For chemical probe selection, find new compounds that inhibit a target using tightly integrated computational methods then optimize and feedback data to data bases and pathways. **(b)** When a target is identified the target based screening workflow can be pursued. **(c)** Phenotypic data is used with integrated computational methods to suggest potential target(s) and optimize ADME properties in parallel, then verify *in vitro*.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 1

Descriptor based QSAR studies

Compound types	Number of molecules in training set	Number of descriptors used	Algorithm used and testing	Refs
Pyrazinoate esters	32	43	Genetic function approximation models, clogP was a key descriptor, the model was tested with 11 external compounds	[71]
N-benzylsalicylthioamides	29	177	2 Multiple Linear Regression (MLR) models for TB with the STATOO program, clogP was a key descriptor, there was no external testing	[72]
Ring substituted-2/4-quinolinecarbaldehyde derivatives	24		PCA analysis, inclusion of logP did not improve model statistics. Actives appeared clustered in a small region of PCA plot	[73]
5-aryl-2-thio1-3,4-oxadiazoles	41	Topological descriptors	Neural networks ($q^2 = 0.8$), not tested externally	[74]
Hydrazides	173	Abraham's descriptors, electronic, geometrical or steric descriptors	MLR subsets were used for modeling. Hydrophobicity could not explain the biological response. For small subsets there were good correlations with test sets ($R^2 > 0.77$)	[75]
Isoniazid derivatives	91	HQSAR and DRAGON descriptors	HQSAR and generated a test set (R^2 0.87) for 24 compounds. The results were better than for PLS-QSAR with 2D descriptors from DRAGON ($R^2 = 0.72$)	[76]
Chalcones and flavonoids	9–33	48	Genetic function approximation, internally cross validated (q^2 0.79–0.94)	[77]

Abbreviations: PCA, principal component analysis; MLR, multiple linear regression; HQSAR, hologram quantitative structure-activity relationship; PLS-QSAR, partial least squares-quantitative structure-activity relationship.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

CoMFA and other 3D-QSAR models

Compound types	Number of molecules in training set	Algorithm used	Statistics	Refs
1,4-dihydropyridines	35	CoMFA and CoMSIA	Cross validated (R^2 of 0.56 and 0.62) as well as external validation ($R^2$0.74 and 0.69)	[78]
Diaryloxymethano-phenanthrene derivatives	37	CoMFA and CoMSIA	CoMFA (q^2 = 0.625) and CoMSIA (q^2 = 0.486) models and 7 compound external test set with very good predictive value	[79]
Deoxythymidine monophosphate derivatives that inhibit thymidine monophosphate	36	molecular field analysis	Alignments performed with least squares (predictive R^2 = 0.70),	[80]
kinase			pharmacophore (0.56) or docked conformations (0.72). Receptor based alignment performed best	
Nitrofuranyl derivatives	95	CoMFA and CoMSIA	Tested with a set of 15 molecules. CoMFA (R^2 = 0.78) outperformed CoMSIA. cLogP and polar surface area or steric bulk did not improve the models	[81]
4-adamantan-1-ylquinoline-2-carboxylic acid alkylidene hydrazides	30	CoMFA and CoMSIA	Models tested with 14 molecules CoMFA (R^2 0.49) and CoMSIA (R^2 0.49)	[82]
Ring substituted quinolines	70	CoMFA and CoMSIA	Tested with 24 molecules. The CoMFA model (R^2 = 0.42). 18 molecules were suggested for synthesis based on the CoMFA predictions.	[83]
Nitroimidazoles	21	Catalyst pharmacophore	Tested with 22 molecules. No test set correlation value reported but correlation looked similar to the training set (R = 0.96)	[84]
1,5-diarylpyrrole derivatives		Catalyst pharmacophore	Had difficulty predicting N-methylpiperazine and thiomorpholine derivatives. Although no numerical prediction data was presented.	[85]

Abbreviations: CoMSIA, comparative molecular similarity indices analysis.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Hybrid methods combining docking and QSAR or pharmacophore methods

Method	Results	Reference
Homology models of DevR, and pharmacophore used to screen 2.5 million compounds, followed by docking with MOE and Gold	Resulted in 11 compounds screened and 4 hits including a phenylcoumarin derivative	[86]
37 enoyl acyl carrier protein reductase carboxamide inhibitors were used to build CoMFA model (tested with 10 compounds $R^2 = 0.88$) followed by the *de novo* molecule design software LEAPFROG	Suggest 13 molecules with improved binding energy values however these have not been synthesized or tested	[87]
29 enoyl acyl carrier protein reductase arylamide inhibitors were used to build CoMFA and CoMSIA models (tested with 8 molecules $R^2 > 0.87$). A pharmacophore was also used to screen the Maybridge database to retrieve 996 hits which were then docked with FlexX.	The CoMFA and CoMSIA scores were used to suggest 20 molecules for future testing.	[88]
Docking and pharmacophore approach used to suggest type II dehydroquinase inhibitors, starting from 45 published inhibitors used to test docking approach and generate GA-MLR QSAR model (35 train, 10 test) using MOE QuaSAR Evolution (q^2 test and train > 0.95). The most active was used for FlexX pharmacophore generation. Also looked at interaction fingerprints.	Predicted 42 active compounds, no test data.	[89]
Combined experimental and computational approach with 12 new imidazoles and triazole derivatives using AUTODOCK to dock molecules in sterol 14α- demethylase followed by free energy of binding calculations.	Good agreement between calculated ΔG_{bind} and experimental data for MIC	[90]
30 5'-thiourea-substituted α- thymidines analogues used to develop receptor independent 4D-QSAR models ($q^2 = 0.83$) for thymidine monophosphate kinase inhibitors. The model was also put into the context of reported crystallographically characterized inhibitor:enzyme interactions	The model was tested with 4 compounds and 3 were predicted within the SD of the assay. Activity also increased with logP.	[91]
31 5'-O-[N-[(Salicyl)sulfamoyl]adenosine inhibitors of MbtA (a salicyl AMP ligase) used with molecular dynamics simulations in a homology model to calculate linear interaction energy (R^2 0.70).	A single validation molecule was predicted with the LIE models to have a K_i of 1.6 nM and the actual value was 0.7 nM	[92]
Docking and molecular dynamics were used to study the binding of the isoniazid metabolite INH-NAD to the enoyl-acyl carrier protein reductase.	Suggested the role of a water molecule in binding. The modeling supported the role of KatG prior to InhA binding	[93]
FlexX and GOLD were used to virtually screen the Chembridge and NCI databases (covering over half a million compounds) against the ATP phosphoribosyl transferase (HisG). Filtering for drug-likeness also used.	50 compounds were tested *in vitro* and 7 were active at 10 μM. Nitrobenzothiazoles were identified as active and co-crystallized, and 19 follow-up compounds found the ChemBridge database (2 of which showed inhibition in the target and whole cell assays)	[94]
UNITY pharmacophore, FlexX docking and structure interaction fingerprint approaches were used to identify compounds in the Maybridge database (59,275 compounds) as potential thymidine monophosphate kinase inhibitors.	10 compounds were ultimately selected and 5 compounds showed MIC < 12.5 μg/ml in whole-cell assays with no cytotoxicity, while the binding of these compounds to enzyme remained to be demonstrated	[43]
CDOCKER used to dock tripeptides into the TB dihydrofolate reductase crystal structure. Molecular dynamics simulation was also performed.	WYY was predicted as potent and selective versus human DHFR. This prediction has yet to be verified	[95]
FlexX used for docking a library of over 19,000 Vichem compounds and Tripos Leadquest compounds into NAD synthetase PknB.	Nine sub-μM inhibitors were found. Additional further docking for NAD kinase inhibitors found that 22 showed activity versus NAD synthetase and one against NAD kinase out of 100 compounds tested	[96]
Catalyst Hypogen pharmacophore and GOLD docking were used to develop the composite model for screening potential thymidine monophosphate kinase inhibitors.	Screened an in-house database of ~500,000 compounds subsequently providing 186 virtual hits that do not appear to have been tested *in vitro*.	[97]
ICM and DOCK were used to virtually screen the University of California, Irvine, ChemDB database and NCI databases to identify AccD5 inhibitors.	One ligand NCI-65828 was found to inhibit AccD5 (an essential acyl-CoA carboxylase carboxyltransferase domain) competitively with an experimental K_i of 13.1 μM.	[98]

Method	Results	Reference
AutoDock used for docking inhibitors to MshB (a GlcNAc-Ins deacetylase).	Docking used to explain mode of binding for inhibitors only.	[99]
AutoDock and GOLD were used to find inhibitors for the adenylation domain of the NAD^+-dependent ligase with bound AMP (LigA).	A novel class of inhibitors, glycosyl ureides, were identified to compete with the NAD^+. Five compounds with docking scores were tested *in vitro* versus LigA, no assessment of correlation.	[100]

Abbreviations: DevR, dormancy regulon; MOE, molecular operating environment; CoMSIA, comparative molecular similarity indices analysis; DHFR, dihydrofolate reductase; AccD5, acyl-CoA carboxylases domain 5; GA-MLR, genetic algorithm-multiple linear regression; KatG, catalase-peroxidase-peroxynitritase; WYY, H-tryptophan- tyrosine-tyrosine-OH.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript