Computational Biology and Chemistry 70 (2017) 65–88

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Virtual screening and repositioning of inconclusive molecules of beta-lactamase Bioassays—A data mining approach

Akshata Gad[a], Andrew Titus Manuel[b], Jinuraj K. R.[c], Lijo John[a], Sajeev R.[a], Shanmuga Priya V. G.[d], Abdul Jaleel U.C.[e,*]

[a] CSIR-OSDD Research Unit, Indian Institute of Science Campus, Bengaluru, Karnataka, 560012, India
[b] Open Source Pharma Foundation, 22-WTC, Brigade Campus, Malleshwaram, Bengaluru, Karnataka, 560055, India
[c] Research and Development Centre, Bharathiar University, Marudhamalai Rd, Coimbatore, Tamil Nadu, 641046, India
[d] Department of Biotechnology, KLE's Dr. M.S.S. College of Engineering and Technology, Belgaum, Karnataka, 590008, India
[e] Open Source Pharma Foundation, 22-WTC, Brigade Campus, Malleshwaram, Bengaluru, Karnataka, 560055, India

ARTICLE INFO

Article history:
Received 20 February 2015
Received in revised form 17 March 2017
Accepted 26 July 2017
Available online 29 July 2017

Keywords:
Inconclusive molecules
PubChem bioassays
Mycobacterium tuberculosis
Self organizing maps
Artificial neural networks

ABSTRACT

This study focuses on the best possible way forward in utilizing inconclusive molecules of PubChem bioassays AID 1332, AID 434987 and AID 434955, which are related to beta-lactamase inhibitors of *Mycobacterium tuberculosis* (Mtb). The inadequacy in the experimental methods that were observed during the *invitro* screening resulted in an inconclusive dataset. This could be due to certain moieties present within the molecules. In order to reconsider such molecules, *insilico* methods can be suggested in place of *invitro* methods For instance, datamining and medicinal chemistry

methods: have been adopted to prioritise the inconclusive dataset into active or inactive molecules. These include the Random Forest algorithm for dataminning, Lilly MedChem rules for virtually screening out the promiscuity, and Self Organizing Maps (SOM) for clustering the active molecules and enlisting them for repositioning through the use of artificial neural networks. These repositioned molecules could then be prioritized for downstream drug discovery analysis.

1. Background

Though large strides have been made in the development of drug regimes, the threat of TB becoming uncontainable continues to pose a challenge to the human race with the emergence of multi-drug-resistant (MDR) and extensively drug-resistant (XDR) strains. The main focus is to shorten the duration of therapy by identifying new inhibitors, and thus, resist the disease and eliminate it successfully. In order to develop novel drugs, novel targets must be determined along with revisiting older ones. Among the various targets, the cell wall based enzymes act as a good source, as the biosynthetic enzymes are absent in mammals (Mdluli and Spigelman, 2006).

One of the major areas that anti-microbial research is focused upon is peptidoglycan biosynthesis. Beta-lactams, a class of antibiotics, critically occludes the pace of synthesis of bacterial cell wall. A significant resistance mechanism towards beta-lactams has been developed in mycobacterium. The four different mechanisms of resistance are inactivation of beta-lactam ring, modification of Penicillin Binding Proteins (PBP), reduction of beta-lactams accessing PBP, and elaboration of antibiotic efflux mechanism (Graumlich, 2004).

Among the various resistance mechanisms, the resistance of Mtb by inactivation of beta-lactam ring is considered to be substantial as beta-lactams hydrolyses the bacterial enzyme beta-lactamases. Thus, the development of beta-lactamase inhibitors, which in turn promotes the beta-lactam activity, is a pivotal point (Madhavan Nampoothiri, 2003).

Penicillins and Cephalosporins are inactivated by beta lactamases, which is expressed by gene BlaC in mycobacterium. Unfortunately, beta-lactamase has never been studied systematically in order to understand how it could be helpful in order to treat TB (Science Daily, 2010). However, Carbapenems and Clavulanate that were used for the sterilization of Mtb cultures indicate that this combination can be used for developing an inhibitor for MDR and XDR infections. The combination therapy of beta-lactams and beta-lactamase inhibitors are not used in Mtb infections as

Abbreviations: MDR, multi-drug-resistant; XDR, extensively drug-resistant; MTb, Mycobacterium tuberculosis; TPR, true positive rate; TP, true positive; FN, false negative; FPR, false positive rate; TN, true negative; FP, false positive; BCR, balance classification rate; ROC, receiver operating characteristic.

* Corresponding author.
E-mail address: jaleel.uc@gmail.com (A.J. U.C.).

http://dx.doi.org/10.1016/j.compbiolchem.2017.07.005
1476-9271/

Clavulanate proves to be a poor inhibitor. All these make it imperative for us to focus upon novel beta-lactamase inhibitors as they could play a vital role in the development of MDR/XDR inhibitors (21st European Congress of Clinical Microbiology and Infectious Diseases, 27th International Congress of Chemotherapy, 2011).

Various databases are available across the web that serve as a resource pool for scientists who are working in the area of neglected diseases (PubChem Bioassay database, CDD database). Among them, PubChem has a large number of TB bioassays, which possess inconclusive datasets along with actives and inactives. These molecules have been concluded as inconclusive for various reasons such as experimental errors, promiscuous nature, property related issues, or bioassay activity outcome. Development of virtual screening activities based on various datamining and artificial intelligence techniques have resulted in the enhancement of research work in Mtb (Ekins et al., 2013, 2010; Ekins and Freundlich, 2013).

In this study, repositioning the inconclusive molecules related to Mtb with a special focus on beta-lactamase bioassays from PubChem were carried out to provide new visual directions for synthetic chemists and medicinal chemists towards the neglected or orphaned molecules.

It is rather impossible to screen several millions of molecules manually. This approach, which has gained tremendous popularity and is an essential constituent of drug discovery research, is feasible by enlisting *insilico* methods. Over the years, several efforts have been created to develop computational models with the help of datamining techniques (Ekins et al., 2015a, 2015b; Litterman et al., 2014). Open source models and crowd sourced collaborative drug discovery activities have benefited with the recent developments in cheminformatics, which in turn has led to the recent developments in the drug discovery industry (Clark et al., 2015a, 2015b; Ekins et al., 2011; Sarker et al., 2012).

Furthermore, virtual screening has overcome the experimental errors of screening the biological compounds, and thus, computational drug design has gained appreciable heights. In the recent past, virtual screening has been considered a dynamic and remunerative concept to explore drugs like compounds/hits. Employing the 3D similarity search or pharmacophore pattern match, ligand based virtual screening grades the novel ligands out of the bioactive conformers. Along with the similarity of various compounds to be screened, 2D or 3D structure of compounds can be found. This method screens the similar compounds against the databases with a common substructure (Reddy et al., 2007).

In order to classify compounds as active or inactive by a model, the model needs to be trained. This can be achieved by making use of bioactive data from high throughput screens and training the learning classifiers of the machine. Discriminant analysis was applied in this study to the virtual screening of bioassay data where the activity of compound was known. Through this approach, predictive models were built to screen the specific bioactive compounds to differentiate them between active and inactive (Schierz, 2009; Periwal and Rajappan, 2011).

Hence, the investigation was to co-relate the molecular properties and descriptor level properties to understand whether it was possible to elucidate a pattern based rule to reposition the inconclusive data, which are related to beta-lactamase bioassays and TB bioassays by various datamining techniques. This could then lead to the prediction of innovative molecular analogues that can be used for further screening and developmental processes.

2. Selection of bioassay to build the model and virtual screening

Data from PubChem bioassay were downloaded for this study. The AID 434987 dataset from PubChem bioassay, a database of NCBI, was used to build the model and was downloaded on 2014-01-20. Since the ratio between active and inactive was less than 1:6, this dataset was chosen to build the model. The dataset that was enlisted for the study as mentioned in PubChem is a screen and counter-screen to identify novel compounds that selectively sensitize Mtb to beta-lactam antibiotics. In addition, the bioassay is of type-confirmatory with total of 1202 tested compounds, out of which 372 compounds are active, 819 compounds are inactive and 11 are inconclusive (Southern Research Specialized Biocontainment Screening Center (TBSL_SelMrpnm), 2010).

The purpose of this analysis was to reposition the inconclusive data of various beta-lactamase inhibitor bioassays of Mtb by virtual screening through datamining methods. Among various PubChem datasets, three bioassay datasets with AIDs 434987, 434955, and 1332 were selected for the study. The data of the selected bioassay either inhibits the beta-lactamase enzymes required for the formation of cell wall of tuberculosis bacterium or inhibits the cell as a whole.

Among the various datasets for Mtb, AID 434987 and AID 434955 datasets were selected and downloaded from PubChem Bioassay developed by Southern Research Institute. These assays possess large compound libraries, with 384-well formatted high throughput screen developed under BSL-3 containment. They were also screened in duplicate in the presence of a sub-lethal (~IC10) dose of the representative beta-lactam antibiotic, Meropenem. Clavulanic acid, an FDA approved drug, was used as a positive control and is a beta-lactamase inhibitor. In the AID 434955 dataset, 328,014 compounds were screened at a concentration of 25 μM in a single dose. Compounds were chosen as actives in AID 434987 and 1202 compounds were selected from the single dose screen with 10 point dose verification in confirmatory assay either in the presence or absence of Meropenem to set up selectivity.

3. Model generation

In ligand based drug discovery, a predictive model is built by using bioassays. The predictive model can either be quantitative or binary. If the model is generated enlisting a binary class of data comprising of active compounds and inactive compounds, it is known as pattern search. Pattern search method was adopted by using a binary class data which was obtained from PubChem. Among the various beta-lactamase bioassays, AID 434987 was selected to build the model for the following reasons.

1. The class ratio between active and inactive compounds have minimum imbalance so that the prediction bias can be reduced.
2. The diversity of the molecule in the model should be high so that the model can predict different classes of molecules with minimum error rate.

3.1. Descriptor generation

Descriptors are various parameters by which a molecule can be fairly represented. The idea of this study was to generate the various descriptors for each molecule and to understand the pattern present in the descriptor data, which was mapped to the active and inactive classes. PowerMV tool was used to generate the descriptors.

3.2. PowerMV

PowerMV is a non-commercial software developed by National Institute of Statistical Sciences for statistical analysis, a descriptor generation, and a similarity search. The software was built for