The emergence of multidrug resistant salmonella biology essay

Science, Biology

ABSTRACTBackground: The emergence of multidrug resistant salmonella enterica serotype typhi in pandemic proportions throughout the world and therefore there is a necessity to speed up the discovery of novel molecules having different modes of action and also less influenced to the resistance formation that would be used as drug for the treatment of salmonellosis particularly typhoid fever. The PhoP regulon is well studied and have now been shown to be critical regulator of number of gene expression whose required for intracellular survival of S. enterica and pathophysiology of disease like typhoid. The evident role of two component PhoP/PhoQ-regulated products in salmonella virulence have motivated attempts to target them as therapeutically. Although high-throughput screens for the inhibitors of the PhoP regulon is available but with faults that they are tedious, expensive as well as time consuming when performed on large scales. And still we need advantageous and expedient method to prioritize the molecules that will be utilized for biological screens which save time and also inexpensive. In this concept insilico methods like Machine Learning are widely applicable technique used to build computational model for highthroughput virtual screens to prioritize molecules for advance study.

Result:

Conclusion:

Background: Salmonella enterica have very wide range of host. There are nearly 2000 different serovars of salmonella genius, among them typhoidal salmonella are S. Typhi and S. paratyphi. Salmonella typhimurium and S. Typhi are most common causative organism for gastrointestinal disease[]. S.

Typhi is extremely infectious for human and animal causes typhoid fever in human which is an acute life threatening infectious disease and there are estimated 33 million cases and 500, 000 deaths worldwide annually[]. Multidrug resistant (MDR) S. Typhi have become a major concern in recent years. MDR typhoid fever increased globally and most of the cases originating from Middle East and Asia especially in the India Pakistan and China []. Therefore discovery of novel anti-typhoidal molecule with less subjective to the resistance formation , has been strong need to control and treatment for the typhoid fever. PhoQ-PhoP two component system regulates the expression of genes which are important to governs virulence and also control the adaptation to host derived defensins and Mg2+ limiting environments in Salmonella enterca and also in several other gram negative bacterial virulence (e. g. Yersinia pestis, Shigella flexneri, Neisseria meningitidis)[]. PhoQ is the inner membrane histidine kinase virulence sensor protein that activated by autophosphorylation at histidine residues in response to number of environmental strait, including low periplasmic Mg2+ level, acidic pH and defensins(rich in phagocytes and small intestinal mucosa) these condition may reflect like intramacrophags and typical human gut during infection. Subsequently cytoplasmic transcriptional regulator protein PhoP activated by accepting phosphate group from PhoQ to promote the transcription of gene which are important for virulence and survival of S. typhi in host. This type of two component signal transduction (TCST) scheme of bacterial expression are not present in human [] and therefore become a novel target for use as therapeutically to developed new antimicrobial drug . With this novel target approach we have consider freely available Bio-assay

data result to generate predictive computational model for drug like molecule with potential to inhibit PhoP regulon. The accomplishment of this was achieved with supervised learning technique by using Machine Learning tools. Furthermore we extend our study to identify common structural entities among the biological active compound towards find out the privileged scaffold.

Methods

Data source

A small molecule screen for inhibitors of the PhoP regulon in Salmonella Typhi, Dataset by high throughput screening method [AID: 1850] was downloaded from Pubchem repository hosted by US National Institutes of health[]. The HTS was conducted in 1536 well black clear bottom microplates with total of 306003 compounds. Compounds that showed greater than 30% inhibition for at least one concentration in the confirmatory PhoP dose response were considered as active compounds. If inhibition at all doses is less than 30%, the compound was defined as " Inactive". In the primary screen the activity cut-off is calculated by the mean of all compound results plus 3x the standard deviation, 33. 5%. And compounds returning value greater than 33. 5% were considered for inclusion in the confirmatory assay. Total 1018 compounds were identified as active, 304870 compound as inactive and 142 compounds labeled as inconclusive from total of 306003 tested compounds. The active and inactive both compounds dataset were downloaded in structure Data format (SDF) file.

Chemical compound descriptor

Chemical descriptors for all compound in both SDF file active and inactive were generated by powerMV[]. PowereMV is a GUI based windows software provides an environment for descriptor generation, visualization, and similarity searching. The large number of compounds were used to study in this bioassay, so the large dataset file was split into smaller SDF file using a Perl script of mayachemtools[]. Total of 179 descriptors were generated for all the compounds in each SDF file(additional file []). These 179 descriptors are based on the two calculation methods , bitstring and continuous. 147 descriptors of Pharmacophore fingerprinting are bitstring based on bioisosteric principals. A " 1" in a bitstring represent the presence of a particular structural feature/fragments and a " 0" its absence. 32 continuous descriptors includes, 24 weight burden number and 8 of properties descriptors. As name implies of weight burden number, inspired by Frank R Burden[] like BCUT by Dr. Pearlman. PowerMV calculation method somewhat different from BCUT. PowerMV uses one of the three properties electronegativity, gasteiger partial charge or atomic lipophilicity and XLogP on the diagonal of the burden connectivity matrix before computing eigen values. The off-diagonal elements were weighted by one of the following values: 2. 5, 5. 0, 7. 5 or 10. 0. Smallest and largest eigen values are used as descriptors. 8 properties descriptors includes XLogP, Polar surface area(PSA), number of rotatable bonds, molecular weight, blood brain indicator (1 indicate the compound pass the BBB, and 0 indicate the compound does not pass the BBB) and bad group indicator(molecule contain toxic group). These properties are helpful for judging the drug like nature of molecules. 179

descriptor generated table for all SDF file were saved in CSV(comma separated values) file format. Then combine CSV file were appended with a extra column as an outcome which represent the bioactivity of compounds and consist of the nominal value 'Active' and 'Inactive'.

Data preprocessing

The combined CSV file loaded in Weka [] and remove the attributes that do not vary at all (i. e. bitstring attributes having all " 0" or " 1" in all the compound throughout the dataset) by weka unsupervised attribute filter. Removing non-informative attributes decreased the extensity of the dataset and improve the efficiency of machine learning. Then dataset was split into 80% training cum validation set and 20% data in independent test set by using of bespoke Perl script .

machine learning classifiers

machine learning is subfield of artificial intelligent, and integrated with different methods and algorithms that educed the functions and rules from large dataset[]. In our dataset we used two methods and three algorithms namely Bayesian method(Naive bayes), decision trees (random forest and J 48)Naive Bayesian classifier based on Bayes theorem[]. It is a very fast computation machine algorithm which is especially important in screening large chemical compound libraries[][]. Bayesian classifier appraises set of attributes(descriptor) to predict the activity (active or inactive) of a compound based on the known attributes of training set of compounds. Each attribute is considered to be statistically independent of all other attributes. Based on a selected attribute, the probability that a compound is active is

proportional to the ratio of active to inactive compounds that have the same value for that attribute and concluding prediction is received by multiplying the descriptor-based probabilities. j48(slightly modified C4. 5 algorithm developed by Ross Quinlan[]) decision tree learning algorithm construct a classification-decision tree for the given data set by recursive partitioning of data[]. Random Forest classifiers uses a number of decision trees, in order to improve the classification rate. Developed by Leo Breiman [] is a symbol of unpruned classification or regression trees made from the random selection of the training data. Random features are selected in the induction process. Prediction is made by aggregating (majority vote for classification or averaging for the regression) the predictions of the ensemble. All above machine learning classification algorithm and data analysis are used in Weka tool kit. Weka is acronym for (Waikato Environment for Knowledge Analysis) developed by researchers at the university of Waikato in New Zealand[]. Weka is java programming based software so it is a platform independent and the collection of many machine learning algorithms, including pre-processing on data(filters), classification, clustering and association rule extraction.

Cost sensitive learning

One of the main problem with the dealing of HTS data is that active compounds are much less in comparison to number of inactive compounds. The minority class have active while the majority class having inactive compounds. In our dataset, (AID1850) have 99. 67% inactive compounds and only 0. 33% active compounds, this is known as imbalanced class problem.

For such problem cost sensitive learning is usually used to trained the model. Because of our interest in correctly classifying the rear class(active compounds as True positive) so cost sensitive learning take place such as misclassification cost into this circumstance. However most of the currently available original classification algorithms do not work well for such cases because they were designed to minimize the overall error rate: the percentage of incorrectly classified of class labels. They ignore the difference between types of the class , means that they're not paying special attention to the active class(TP). This implicitly assumes that all misclassification errors are equally costly[] but this is far from this dataset. There are mainly two techniques in use to overcome this problem of highly imbalanced data. One is the Direct method that is to design independent cost sensitive classifier algorithm such as ICET(Turney, 1995)[] and cost sensitive decision trees by Ling et al.[]. Second one is Meta learning which converts the cost insensitive error-based classifier into cost sensitive ones such as CSC-CostSensitiveClassifier(Witten and Frank, 2005)[] and MetaCost by Pedaro Domingos, 1999.[]. In this method introduces a bias(i. e., in our case set a high cost to miss classification of active compound-(FN)) by cost metrics C(i, j) in training dataset and trying to minimize the overall cost, where i is the actual class and J is the predicted class. Meta learning has been used in our experiment. In Weka, we used CSC for Naive Bayes, Random forest and MetaCost has used for J48. We have used default option for Naive bayes and Random Forest but for j48, unpruned option was set to true in Weka. In prior research, shown that these settings are optimal and gives the better result with their corresponding classifier.[1]. A positive aspects of Meta cost is that

is uses bagging on the decision trees to get trustworthy probability calculation of training data set and then it is used to relabel the training dataset to build an ensemble classifier(cost sensitive).[2]. A bad thing about Meta cost is that it takes longer time to run in comparison to CSC. Cost matrix may be seen as a layout of confusion metrics. For binary class problem a 2x2 Weka cost metrics has four sections, True positive(TP)- in our experiment active compound correctly predicted as Active; False-positive(FP)- Inactive compounds incorrectly predicted as Active; True negative(TN)- Inactive compound correctly predicted as Inactive, and False-negative(FN)- active compound incorrectly classifies as inactive. In our case, if true positive classified as false negative(FN) that means PhoP regulon inhibitor compound misclassifies as Inactive that is more expensive in comparison to non inhibitor compound predicted as inhibitor compound(FP). Therefore FN is more important, misclassification cost have been set for False-negative. Increasing the misclassification cost of False-negative will potentially increase the True positive and False positive both. For maintaining generality we limit the maximum false positive rate ≤ 20%.

Cross-Validation

Cross validation option was used with 5 fold for evaluation the model in training session. Cross validation is a technique which applied for evaluation of the model prior to independent dataset test. In N fold cross validation the training dataset is fragmented into N number of subset. Then(N-(N-1)) subset is extracted for test and remaining are used for training. And finally model

was tested with independent 20% test dataset which compounds yet completely unknown to the model.

Model performance parameters

To evaluate the model performance, we used various statistical performance measures that include Accuracy(Ac), Sensitivity(Sn), Specificity(Sp), BCR, MCC and ROC. Accuracy is the overall percentage of correct prediction of PhoP regulon inhibitor and non-inhibitor compounds. Sensitivity is the proportion of correctly predicted inhibitor compounds as active. Specificity is the percentage of non-inhibitor compounds predicted as inactive. Balanced classifier rate(BCR) is the mean of Sn and Sp and this value introduces a balance in the classification of unbalanced dataset. Matthew's Correlation Coefficient (MCC) is the fitness function for model optimization. The value of MCC considered between the range(-1 to +1); 1 is regarded as a perfect prediction and 0 is regarded as a random prediction. Receiver operating characteristic(ROC) plot is the graphical curve between true positive rate and false positive rate that evaluate the performance of classifier by use of AUC value; Area under the curve(AUC) value is the probability that classifier gives higher score to active compound in comparison to inactive compound in case of randomly chosen compounds. Ac = Sn = Sp = MCC = BCR = Where TP and TN correctly predicted as active and inactive compounds; FP and FN are falsely predicted active and inactive compounds.

Maximum common substructure search

As per the identification of potentially enriched active molecules in chemical compound, we go for maximum common substructure(MCS) approach. 2D

active and inactive datasets were converted into 3-D active and inactive datasets by using command line program MolConverter available in JChem[] and for the identification of substructure which are common to a pair of molecule, we used hierarchical clustering algorithm " LibMCS" which is available from ChemAxon[]. The classification of the molecules by the LibMCS program, generate different cluster with an approach that structures which share a large common substructure are grouped together and represent as a hierarchal dendrogram. The size of MCS is a functionality of the numbers of the constituent atoms which empirically set to a threshold of 10 for the better clustering. generated molecular scaffold via clustering was used as SMILES query to search for substructure in both target, active and inactive datasets by using " Jcsearch" algorithm available in ChemAxon. The evaluation of substructure enrichment and their significance were done by statistical method. Chi-square test evaluate the enrichment and P values were used to evaluate their significance. To improve the culmination we filter out indigent scaffold which has less than 1% of matches in active dataset and also calculated enrichment factor and remove the scaffold that have not empirical threshold of ≤ 2.

Result and discussion

confirmatory in nature bioassay dataset(AID1850) downloaded from Pubchem repository were used to calculate 179 molecular descriptor(Additional file-1) using PowerMV. 25 number of useless molecular descriptor(Additional file-2) removed in data processing as described in methodology section and finally remaining 154 molecular

descriptor(Additional file-3) further used in classification task. initially standard base classifier were used to develop models but having low True Positive(TP) due to biased nature of standard classifier towards majority class(inactive compound). Therefore cost sensitive learning were used : as we rise the False negative(FN) cost , TP rate goes up, with slightly decrement in True negative(TN) rate. Owing the decrement in actual negative instances, we maintained the reasonable specificity with not more than our threshold limit of FP rate 20%. Thus number of models were trained using different cost of FN. Final misclassification cost of FN (Table no-1) was selected for the models of each classifier by evaluating model performance with different statistical performance measurement. Best selected among the trained model of each classifier, performance statistics are shown in Table no-2, these result are based on testing of 20% independent dataset. Table no.-1 data shows, bayesian theorem based Naive Bayes classifier required a lower misclassification cost in comparison to Random Forest and j48, while maintaining FP rate below to our threshold limit (e. i. 20%). In our experiment Naive Bayes was very quick in building the models. Owing the extremely imbalanced data classification, the overall classification accuracy alone may not turn out to be an appropriate measure performance. Therefore another classification measure, Balanced classification Rate also known as Balanced Accuracy was calculated that introduced a balance in the classification by providing an average of sensitivity and specificity. As can be observed from Figure-1, BCR value is optimum for Random Forest among other classifier. Evaluating the relative classifier performance by using area under the curve (AUC) value , obtained from plotting the ROC curve between

TP rate and FP rate. Random Forest covers the maximum AUC as compared to other classifier(see figure-2). In order to vindicate the classifier's predictive potentiality of actual biological activity (active or inactive) of compounds, we have used to measure sensitivity and specificity. As figure-3 shows, all classifier are more than 80% specific in their prediction while the sensitivity is higher for Random Forest followed by J48 and Naive Bayes.

Evaluation of substructure

In order to find out the most active substructures among the active dataset LIBMCS tool was used , which resulted in 1285 most common substructures from a set of 1018 active compounds. By taking the minimum class size 10 i. e the basis of clustering was to group at least 10 similar atoms together, at top level count 6 the total cluster count was 188 in which 76 singletons were removed which showed no similarity during clustering. Finally the most active scaffolds were selected whose frequency was > 1% of actives, further our analysis was validated by taking into parameters of chi square test, p value less than . 01 and enrichment factor > 2(which correspond to the ratio of frequency in active to inactive). List of most active scaffolds (14) are shown in Table no-3ClassifiercostRandom forest85991J481176Naive Bayes210

TABLE 1 Misclassification cost used for False Negatives with each classifier.

ClassifierTP%FP%TN%FN%ROCAccuracyMCCBCRSensitivitySpecificityNaive Bayes

66. 5

19. 9

80. 1

33. 5

80. 2

80. 08

0. 067

73. 32

66. 50

80. 13

Random Forest

87. 7

18. 5

81. 5

12. 3

91. 5

81. 52

0. 1

84. 59

87. 68

81. 49

J48

78. 3

19. 6

80. 4

21. 7

80. 9

80. 37

0. 085

79. 35

78. 33

80. 37

TABLE 2 Statistics of best predictive models generated by three different classifier.

Figure 1 Comparative analysis of Accuracy & BCR

Figure 2 Plot of sensitivity and specificity

Figure 3 Roc Plot- shows the significant AUC value for naive bayes, random forest and j48

S. no.

Smiles

Matches in Actives

Matches in Inactives

Chi-square

p-value

Enrichment Factor

1= JCSYSStructure(" 09C6824E98B8C6DD8ED4C62C8AFD607F")15461082. 350. 0097. 662= JCSYSStructure(" F3E68FBD5D0EA5D203A701E569B99385")17103692. 630. 0049. 433= JCSYSStructure(" AFA3B3613C0AB2530CC3FC515FFE731C")19135669. 450. 0042. 154= JCSYSStructure(" 66B561336B982F61E53E6A1061880BCE")14109452. 910. 0038. 475= JCSYSStructure(" F072A7B2E9A6F3EA2823A4B18D53A2E9")37804420. 480. 0013. 786= JCSYSStructure(" 47FC1A8444F37C3A9392D6936850E38D")471216439. 010. 0011. 567= JCSYSStructure(" CB89F767A05550B655FF62DA25F1632A")16448135. 980. 0010. 70

Table 3 Potentially enriched scaffold in active dataset.