

OPEN ACCESS

Full open access to this and thousands of other papers at <http://www.la-press.com>.

Support Vector Machine Based Classification Model for Screening *Plasmodium falciparum* Proliferation Inhibitors and Non-Inhibitors

Sangeetha Subramaniam¹, Monica Mehrotra² and Dinesh Gupta¹

¹Bioinformatics Laboratory, Structural and Computational Biology Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi, India. ²Department of Computer Science, Jamia Millia Islamia, New Delhi, India. Corresponding author email: dinesh@icgeb.res.in

Abstract: There is an urgent need to develop novel anti-malarials in view of the increasing disease burden and growing resistance of the currently used drugs against the malarial parasites. Proliferation inhibitors targeting *P. falciparum* intraerythrocytic cycle are one of the important classes of compounds being explored for its potential to be novel anti-malarials. Support Vector Machine (SVM) based model developed by us can facilitate rapid screening of large and diverse chemical libraries by reducing false hits and prioritising compounds before setting up expensive High Throughput Screening experiment. The SVM model, trained with molecular descriptors of proliferation inhibitors and non-inhibitors, displayed a satisfactory performance on cross validations and independent data set, with an average accuracy of 83% and AUC of 0.88. Intriguingly, the method displayed remarkable accuracy for the recently submitted *P. falciparum* whole cell screening datasets. The method also predicted several inhibitors in the National Cancer Institute diversity set, mostly similar to the known inhibitors.

Keywords: Plasmodium, proliferation, inhibitor, support vector machine, screening, in silico, drug, molecular descriptors

Biomedical Engineering and Computational Biology 2011:3 13–24

doi: [10.4137/BECB.S7503](https://doi.org/10.4137/BECB.S7503)

This article is available from <http://www.la-press.com>.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.



Introduction

Malaria is a devastating disease causing millions of death annually, apart from thousands of man hours lost to morbidity.¹ The majority of deaths due to malaria are caused by *P. falciparum*, the most virulent amongst the rest of the species that cause the disease. The mounting resistance and failure of existing first-line antimalarial drugs has exacerbated the condition leading to an urgent need to develop novel anti-malarials.

Amongst various experimental methods, the experimental cell based assays to identify growth inhibitors of *P. falciparum* has been one of the promising approaches for novel antimalarial drug discovery. The technique has shown success in identifying several novel chemical scaffolds with antimalarial activity.^{2,3} Cell based bioassays make use of living organisms, enabling the simultaneous testing of all drug targets for their viability in the presence of the test compounds. Identification of *P. falciparum* intraerythrocytic cycle proliferation inhibitors has gained much attention as many of these compounds have successfully inhibited the parasite growth at a very low concentration.^{4,5} Development of in silico models for prediction of proliferation inhibitors against *P. falciparum* will aid research experiments aimed towards identification of novel antimalarial leads. The study reported here focuses on development of a SVM based classification method for *P. falciparum* proliferation inhibitors and non-inhibitors.

The striking growth and complexity of High Throughput Screening (HTS) data has increased the importance of data mining techniques to aid efficient data analysis and decision-making at crucial phase of drug discovery.⁶ Such techniques are often helpful to discover meaningful patterns and rules in the screened data. These patterns form the basis for building models that are effectively applied to prioritize compounds for the subsequent phases. Data mining methods can assist identification of false leads at an early stage and also facilitate understanding of Structural and Activity Relationships.⁷ Supervised and unsupervised methods are increasingly being applied to build predictive bioactivity compound models.⁸ In various studies, classification of compounds has been carried out using machine learning methods like Decision Tree (DT), k-Nearest Neighbours (kNN), Artificial Neural Networks (ANN),

PLS Discriminant Analysis (PLS-DA) and all of them have shown statistically significant performance.⁹ It is encouraging to note that the existing mathematical methods in Quantitative Structural and Activity Relationship (QSAR) field are being constantly upgraded and novel mathematical algorithms are continuously evolving. At the same time, the increasing availability of published compounds assays in PubChem database has stimulated greater interest to apply these robust methods, leading to development of highly accurate predictive models.^{10–12}

In recent years, SVM based classification has gained wide usage in Ligand Based Virtual Screening (LBVS) mainly due to its efficient generalization capabilities and empirical performance.^{13–15} SVM based ligand screening has been illustrated as an ideal tool for rapid screening of large compound libraries with enhanced hit rate and better coverage.^{16,17} Unlike most of the LBVS methods which work on similarity based principles, SVM based classification has been shown to yield structurally diverse hits.¹⁸ In a comparative study conducted by Plewczynski et al.¹⁹ SVM with a linear kernel was found to be the best performing algorithm, compared to the other methods namely; kNN, ANN, DT, Random Forest (RF) and Naïve Bayesian Classification (NBC). Summarily, the performance of SVM methods is better when compared with above-discussed methods and hence we have used SVM in our studies. In previous studies, linear methods have been applied for classification of antimalarial compounds; however, there are fewer reports about usage of non-linear methods.^{20,21} In this study, we have developed linear as well as non-linear SVM models to classify compounds for anti-proliferative activity against *P. falciparum*.

Materials and Methods

Generation of training and independent testing set

All the molecular structures for generating SVM models were retrieved from the PubChem bioassay data corresponding to the bioassay ID “AID-1815”.^{4,5} The Bioassay reports 441 active compounds with the potency ranging from 0.06 μM to 14.12 μM . The assay was based on qHTS for differential inhibitors of proliferation of *P. falciparum* line 7G8, derived from a malarial isolate from Brazil.



Further assay details are available at the PubChem bioassay database website (<http://pubchem.ncbi.nlm.nih.gov/>).

In preparing the training and independent test dataset, we have considered all the compounds labelled as active or inactive for their inclusion in the positive and negative training dataset. In the bioassay, 441 compounds are reported as active (potency ranging between 0.06 μM to 14.12 μM) and 558 compounds as inactive, ie, a total of 999 compounds. These compounds were pre-processed for removal of redundant compounds, which resulted in 426 active and 533 inactive compounds. All the non-redundant compounds, 959 in numbers, were standardized and hydrogen atoms were added using JChem 5.2.²² The data set of 959 compounds was divided into training and testing set such that exactly 80% was reserved for training and the remaining 20% was retained for independent testing, ie, not to be included in the training step. Thus the training data comprised of total 640 compounds, while the test set comprised of 319 compounds (Table 1). SVM models described here after are developed based on the training set of 290 active compounds and 350 inactive compounds. Some of the potent proliferation inhibitors in the training data are shown in Table 2. The training and testing dataset is available online as supplementary material (Supplementary file 1 and 2).

Descriptor calculation and selection

Molecular descriptors are the numeric representation of physico-chemical features extracted from various structural representation of a molecular structure.²³ Such a quantitative representation is obtained as the result of a logical and mathematical procedure that transforms chemical information encoded within a symbolic representation of a molecule into a useful number. In this work, a number of 0D (constitutional descriptors), 1D (functional group counts), 2D (topological, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency matrices, Burden Eigen values,

topological charge indices, Eigen-value based indices) and 3D (Randic molecular profiles, geometric, RDF, 3D-Morse, WHIM, GETWAY) descriptors were calculated using DRAGON software.²⁴ Details of individual descriptors can be found in the reference manual of DRAGON software. The list of descriptors used in the study for developing different SVM models is presented in Table 3. Calculations for 0D, 1D and 2D descriptors were based on 2D structures of the compounds where as 3D descriptors calculations were based on JChem generated single low energy conformers. We generated three models based on different sets of descriptors; the first one based on 0D, 1D and 2D descriptors, the second one based exclusively on 2D descriptors and the third model based exclusively on 3D descriptors. The total number of selected descriptors was above 300 in each case. In order to reduce redundancy and noise in the training data, we reduced the number of descriptors in each case. For instance, in case of the model based on 0D, 1D and 2D descriptors, the total number of calculated descriptors was 383. We reduced the total number of descriptors to 184 by the following approach: firstly, descriptors with the same values and near-constant descriptors were eliminated. Secondly, redundant descriptors were removed by pair correlation method. The pair wise correlations for all descriptors were examined and one of the two descriptors with the correlation coefficient r of 0.9 and higher was excluded. Finally, three different models were developed using 184 (0D, 1D, 2D), 112 (2D) and 195 (3D) descriptors respectively. SVM training and testing files require normalized data input, hence we normalized our training data to range within -1 to $+1$.

SVM algorithm

The SVM method was developed by Vapnik.²⁵ SVM algorithms project input data into a high-dimensional feature space using kernel functions, so that an optimal plane (maximal-margin hyper plane) may be drawn which can demarcate positive and negative datasets. The hyper plane is dependent on choice of kernel function and representative training examples, called support vectors. Optimized SVM classification model is generated by iterations of learning and evaluations, based

Table 1. Compound dataset used in this study.

Data set	Inhibitors	Non-inhibitors	Total
Training	290	350	640
Test	136	183	319

Table 2. Selected proliferation inhibitors in the training dataset.

Molecule	Structure	Molecular weight	Potency [μM]
Dequalinium		527.571	0.0891
Quinacrine		472.879	0.1778
Clotrimazole		344.837	0.2512
Reserpine		608.679	0.2818

on optimized choice of training support vectors, kernel functions and parameters. In developing SVM models for this work, we have used LIBSVM, available freely at (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>).²⁶ The SVM model built in

this study is based on C-SVC (C-Support Vector Classification) algorithm implementation of LIBSVM. As the number of features used here is less than the number of instances, we primarily used a non-linear kernel for building the SVM model.



However, models based on linear kernel were also developed to compare their performance. A coarse grid-based optimization of the kernel parameters C and the hyper parameter γ was performed to achieve the highest classification accuracy.

Model validation

The training data set of 640 compounds was subjected to five-fold cross validation to find the best kernel parameters C and γ by maximizing the accuracy and minimizing the error. In five-fold cross-validations, the training data is split into 5 folds; one fold is used for testing, the remaining ones for training. This is iterated five times, such that each of the data sets is used as a test data. The optimum values of C and γ were then used to retrain the SVM model. The performance of the models was also assessed on the independent test dataset, using standard statistical measures namely- sensitivity: the percentage of correctly predicted active compounds, specificity: the percentage of correctly predicted inactive compounds, accuracy: the percentage of correctly predicted active and inactive compounds. In addition, balanced measures like MCC, Balanced Accuracy (BAC) and AUC (Area under ROC curve) were also computed.²⁷ MCC = 1 indicates a perfect prediction while MCC = 0 indicates a random prediction.

These evaluation measures can be mathematically expressed as:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

$$\text{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

$$\text{BAC} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives and FP is the number of false positives.

Principal Component Analysis (PCA) and Applicability Domain (AD)

As expected, the models based on machine learning methods normally show good performance for compounds that share similar properties as those in the training set. Thus, it is of ever-increasing concern to define the AD of the models, and to check if it is valid for any new molecules. AD is the boundary defined by the descriptor space in the training data. Any new chemical compound should essentially be positioned in the boundary of the chemical space of the training set, in order to be qualified for reliable prediction.²⁸ Several simple and complex approaches are used to define AD; based on range, distance, geometric and density distribution. One of the simplest and widely applied approach is the AD based on range-based definition with a preliminary PCA rotation.²⁹ In the present study, we have defined the AD of the model and evaluated its validity on the test set and the screening dataset based on Principal Component (PC) ranges. This method will be helpful to confirm whether a new compound is inside or outside the AD. PCA based definition of AD, reduces the higher dimensionality of the data (due to large number of descriptors) and facilitates simple exploration besides maintaining the variation of the data. This is achieved by identifying directions, or PCs, along which there is maximal variation in the data. Each PC is expressed as a linear combination of the original descriptors. It may be noted that PCs are orthogonal to each other, and the correlation between any two PCs is zero. PCA in the study was performed using R package.³⁰ The PCA was carried out for the training data of the best model that showed highest classification accuracy. PCA of the independent test set and screening set of NCI diversity set II was also performed in order to validate the applications of the model.

SVM model as virtual screening tool

We used NCI diversity set of 1364 compounds retrieved from the NCI/DTP Open Chemical Repository (http://dtp.nci.nih.gov/branches/dscb/div2_explanation.html) for virtual screening purpose.³¹ The compounds were processed in the same way as done for the training set and descriptors were calculated. Subsequently, 1328 compounds were suitable for



descriptor calculation and were predicted for their activity by the best SVM model.

Results and Discussion

In the present study, we have developed SVM based model for prediction of proliferation inhibitors of *P. falciparum* in erythrocytes based on the bioassay results from PubChem bioassay ID “AID 1815”. SVM models were generated using three different sets of descriptors; the first model was based on descriptors belonging to 0D, 1D and 2D category, the second model was based on 2D descriptors and the third model was based on 3D descriptors (Table 3). Non-linear models based on Radial Basis Function (RBF) kernel and linear models were developed for each category of the descriptors. A five-fold cross validation method was used to select the best kernel parameters and to evaluate the self-consistency of the data set in each case. The performance of the models was assessed using a test dataset of 319 compounds, not used in the training process.

Model validation

The overall cross validation accuracy of the models is in the range of 80% to 83%, this suggests the self-consistency of the data and also validates the reliability of the models (Table 4). Area Under the ROC curve (AUC) values for all the models (~0.88) indicates an overall good performance of the models than random classification.

Performance of model based on 0D, 1D and 2D descriptors

Best kernel parameters determined by five-fold cross validation and the corresponding results

obtained with each model are illustrated in Table 4. The model was based on a set of 184 descriptors belonging to 0D, 1D and 2D category which yielded highest accuracy in cross validation as well as over independent test set. The model performed consistently well with an accuracy of 83%, and an AUC of 0.88 in five-fold cross validation. The model was able to correctly classify 117 inhibitors (86%) and 160 non-inhibitors (87%) with an overall accuracy of 87% and MCC of 0.73. Although the number of inactive compounds is slightly higher than the active compounds in the training set, almost equal sensitivity and specificity was obtained. This signifies the balanced performance of the model with respect to good recognition rate and low false prediction rate. The overall performance of the model was found to be satisfactory as evident from the independent testing data performance. The better accuracy of the model can be attributed to the appropriate choice of 0D, 1D and 2D descriptors that were capable to discriminate proliferation inhibitors and non-inhibitors. The model based on these descriptors showed consistent and optimum performance when compared with other models.

Performance of models based on 2D and 3D descriptors

The non-linear model based on 2D descriptors showed second best performance with a five-fold cross validation accuracy of 82%, overall testing accuracy of 85%, sensitivity of 84% and MCC measure of 0.69. Some of the 2D descriptors applied in this model are overlapping with the ones used in the first model. This implies the specific contribution of 2D descriptors in better discrimina-

Table 3. Molecular descriptors used in the development of SVM models.

Classifier	Descriptor category	Descriptor class	Total number of descriptors
1	0D, 1D, 2D	Constitutional, topological, connectivity indices, functional group counts, molecular properties	184
2	2D	Topological, Walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency matrices, Burden Eigen values, topological charge indices, Eigen-value based indices	112
3	3D	Randic molecular profiles, geometric descriptors, RDF descriptors, 3D-Morse descriptors, WHIM descriptors, GETWAY descriptors	195

Table 4. SVM model parameters and evaluation of classification performance.

Classifier	Training			Testing						
	Kernel	Parameters	Cross validation accuracy	AUC	Overall Accuracy	Sensitivity	Specificity	MCC	Precision	BAC
1	RBF	C = 32 G = 0.0078	83%	0.88	87%	86%	87%	0.73	84%	0.87
	Linear	C = 0.125	82%	0.88	84%	81%	87%	0.68	82%	0.84
2	RBF	C = 8 G = 0.125	82%	0.87	85%	84%	85%	0.69	80%	0.84
	Linear	C = 0.125	79%	0.87	82%	84%	80%	0.64	76%	0.82
3	RBF	C = 128 G = 0.0019	80%	0.88	81%	79%	82%	0.62	78%	0.81
	Linear	C = 0.5	80%	0.88	80%	79%	81%	0.60	76%	0.80

Abbreviations: AUC, Area under Curve; MCC, Matthews Correlation Coefficient; BAC, Balanced Accuracy.

tion of active and inactive compounds with good sensitivity and specificity than models based on 3D descriptors. Linear model based on 2D descriptors had the lowest cross validation accuracy although with good testing accuracy (81%) comparable to other models. As shown in Table 4, the model based on 3D descriptors ranks last in the testing accuracy, specificity and sensitivity. Perhaps, the overall limited structural diversity in the compounds could be a limiting step for performance of shape based 3D descriptors. In general, all the models showed a balance in terms of their specificity and sensitivity as demonstrated by the BAC. The overall BAC of all SVM models ranges from 0.80 to 0.87. However, in all the cases RBF based models outperformed corresponding linear models with higher classification accuracy.

Applicability Domain

PCA was applied here to define the AD of the best model and also to map the active and inactive compounds in their respective chemical spaces. PCs are basically the linear combinations of the original 184 descriptors used in this study. The AD is calculated on the basis of the PC ranges. The minimum and maximum values of principal components are set by considering all the compounds in the training data set. Figure 1 shows the first three principal components of the compounds in the training set that has been used to define the AD of the model. The compounds in the independent testing set were also found to be within the AD (Fig. 2). PCA results

reveal in general that, the active and inactive compounds occupy different clusters in the chemical space, although there was no clear boundary between the two classes. The training data shows limited structural diversity, which poses a restraint on the sensitivity and specificity of the model. These parameters could be apparently improved by increasing the number of diverse structures in training set.

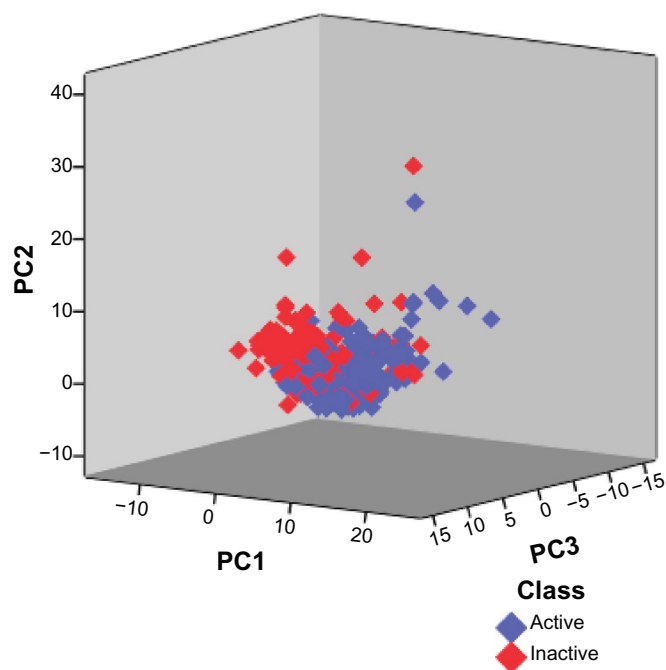


Figure 1. Visualization of chemical space in training dataset. Proliferation inhibitors (blue diamonds) and non-inhibitors (red diamonds) are represented using the first three Principal Components. The figure depicts the range of Principal Components of the compounds in the training set that define the applicability domain (AD).

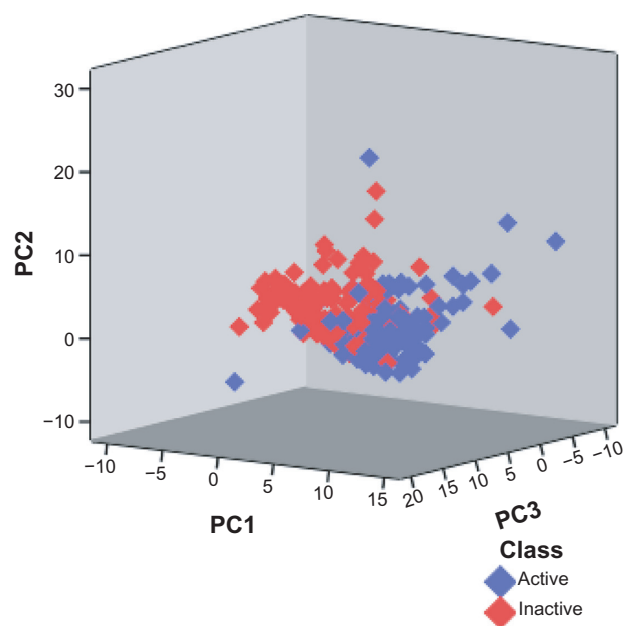


Figure 2. Visualization of chemical space in testing dataset. The figure illustrates the compounds in the independent testing dataset lying within the applicability domain of the classifier.

Additional validation using ChEMBL-NTD datasets

While we were developing the models, three novel datasets of proliferation inhibitors of *P. falciparum* were submitted to the ChEMBL-Neglected Tropical Disease database (www.ebi.ac.uk/chemblntd) from three sources namely: GSK TCAMS Dataset, Novartis-GNF Malaria Box Dataset and St. Jude Children's Research Hospital Dataset.^{32–35} We used the datasets to perform additional testing on the model developed by us. Prior to screening, we ensured that none of the compounds in these datasets overlaps with those in the training dataset. Two such overlapping entries found in Novartis-GNF Malaria Box Dataset were removed before screening. The performance of the SVM model over these datasets is shown in Table 5. The SVM model correctly

predicted 89% (12082/13519), 83% (4750/5692), and 90% (1384/1535) of the experimentally verified inhibitors in the GSK, Novartis and St. Jude's datasets respectively. These results suggest that the SVM model is equally effective in identifying potential hits in virtual screening of large libraries with reasonable AD.

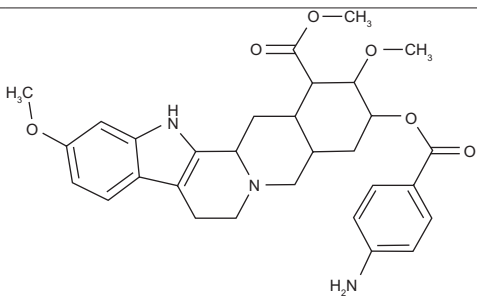
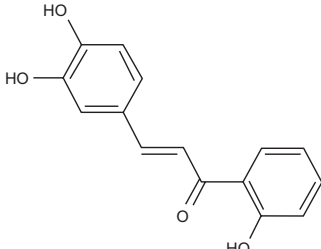
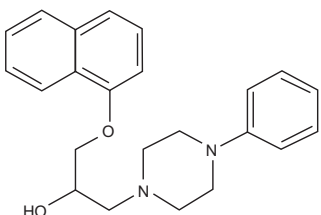
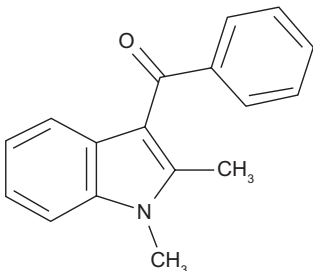
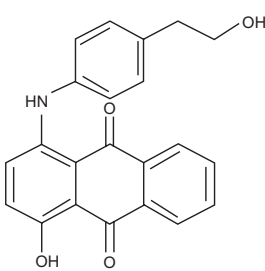
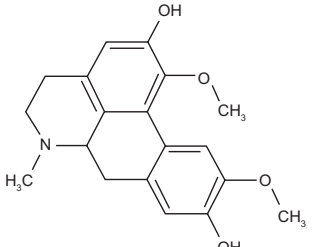
Virtual screening of inhibitors

We have utilised the best SVM model based on 0D, 1D and 2D descriptor category for identifying further novel inhibitors from NCI diversity collection of 1364 compounds. Only 1328 compounds passed through the descriptor calculation. First we tested, if all 1328 compounds were within the AD of the model using the first three principal component ranges of the 184 descriptors (as described in the methods). About 70 compounds violated the descriptor ranges observed for compounds in the training set. Therefore we considered them unreliable for prediction. The outliers were discarded and the remaining 1257 compounds were predicted using best SVM model. The model predicted about 580 compounds as positive and remaining 677 as negative. In the NCI diversity set, we observed that there were four known proliferation inhibitors which were correctly classified. The predicted compounds were prioritised according to the probability score of LIBSVM. Further, we compared predicted active compounds to those in the training dataset, to check their similarity in terms of Tanimoto coefficient. The Tanimoto coefficient for the 580 predicted positive NCI diversity set compounds against the 290 positive training compounds ranged from 0.98 to 0.24. Some of the predicted active compounds and their corresponding maximum Tanimoto score to the compounds in the training data are shown in Table 6.

Table 5. Performance of the SVM model in validating ChEMBL-NTD datasets.

Dataset	Number of active compounds in the dataset	Number of compounds correctly identified	Percentage of true hits
GSK TCAMS dataset	13519	12082	89%
Novartis-GNF malaria box dataset	5692	4750	83%
St. Jude children's research hospital dataset	1384	1535	90%

**Table 6.** Selected virtual hits from NCI diverse set collection.

Molecule	Structure	Molecular weight	Tanimoto score
Methyl 18-((4-aminobenzoyl)oxy)-11,17-dimethoxy-yohimban-16-carboxylate		533.623	0.95
3-(3,4-dihydroxyphenyl)-1-(2-hydroxyphenyl)-2-propen-1-one		256.257	0.86
1-(1-naphthyloxy)-3-(4-phenyl-1-piperazinyl)-2-propanol		362.471	0.91
(1,2-dimethyl-1H-indol-3-yl)(phenyl)methanone		249.312	0.90
1-hydroxy-4-(4-(2-hydroxyethyl)anilino)anthra-9,10-quinone		359.381	0.80
1,10-dimethoxy-6-methyl-5,6,6a,7-tetrahydro-4H-dibenzo[de,g]quinoline-2,9-diol		327.379	0.95

(Continued)

Table 6. (Continued)

Molecule	Structure	Molecular weight	Tanimoto score
methyl 6-methyl-9,10-didehydroergoline-8-carboxylate		282.341	0.88
4-(2-(6-quinolinyl)vinyl)aniline		246.311	0.58
1-(3-chlorophenyl)-3-ethyl-5-(2-phenylvinyl)-1,2,3,4-tetrahydropyrimido[5,4-c]quinolin-9-yl methyl ether		455.985	0.51

Conclusion

The SVM model based on 184 0D, 1D and 2D descriptors of the inhibitors exhibited the highest accuracy with lower false-hit rate. The selected molecular descriptors have sufficiently captured the features required to discriminate *P. falciparum* intra erythrocytic cycle proliferation inhibitors from non-inhibitors. The predictive power of the optimized model with good performance on three additional validation (ChEMBL-NTD) datasets indicates that it can be equally effective in selecting potential hits in screening large libraries. Several new compounds predicted as inhibitors from NCI diverse set have shown good similarity to the known proliferation inhibitors. The SVM model developed in this study is fast and precise enough to be applied for large scale screening of proliferation inhibitors of *P. falciparum*. The large repositories of chemical compounds, for example PubChem and ChEMBL-NTD can be a rich source for generating such

quality predictive models. Efforts to publish more bioassays for neglected diseases like malaria would benefit from such data mining techniques to support decision-making in drug discovery. However the generalization capability of a model largely depends on the quality and the diversity of the data set. Such models can be ideal for quick screening of potential bioactive molecules from large chemical libraries and facilitate lead identification.

List of Abbreviations

ANN, Artificial Neural Network; AD, Applicability Domain; AUC, Area Under Curve; C-SVC, C-Support Vector Classification; DT, Decision Tree; HTS, High Throughput screening; kNN, k-Nearest-Neighbor; LBVS, Ligand Based Virtual Screening; MCC, Matthews Correlation Coefficient; MDDR, MDL Drug Data Report; NBC, Naïve Bayesian Classification; PCA, Principal Component Analysis; QSAR, Quantitative and Structural Activity Relationship;



RBF, Radial Basis Function; ROC, Receiver Operating Characteristic; SVM, Support Vector Machine.

Acknowledgements

Department of Biotechnology (DBT, India) grant for “Bioinformatics Infrastructure Facility” at ICGB and Indian Council of Medical Research (ICMR) fellowship to Sangeetha Subramaniam is duly acknowledged.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. World Health Organization (World malaria report). <http://www.who.int/malaria/publications/atoz/9789241563901/en/index.html>. 2009.
2. Baniecki ML, Wirth DF, Clardy J. High-Throughput Plasmodium falciparum Growth Assay for Malaria Drug Discovery. *Antimicrob Agents Chemother*. February 1, 2007;51(2):716–23.
3. Jennifer LW, Ally PL, Anang AS, Fred EC, Guy RK, Joseph LD. Searching for New Antimalarial Therapeutics amongst Known Drugs. *Chemical Biology and Drug Design*. 2006;67(6):409–16.
4. Plouffe D, Brinker A, McNamara C, et al. In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proceedings of the National Academy of Sciences of the United States of America*. July 1, 2008;105(26):9059–64.
5. qHTS for differential inhibitors of proliferation of Plasmodium falciparum line 7G8. <http://pubchem.ncbi.nlm.nih.gov/>. Accessed November 12, 2009.
6. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical practice in high-throughput screening data analysis. *Nat Biotech*. 2006;24(2):167–75.
7. Harper G, Pickett SD. Methods for mining HTS data. *Drug Discov Today*. 2006;11(15–16):694–9.
8. Muegge I, Oloff S. Advances in virtual screening. *Drug Discovery Today: Technologies*. 2006;3(4):405–11.
9. Melville JL, Burke EK, Hirst JD. Machine learning in virtual screening. *Combinatorial Chemistry and High Throughput Screening*. 2009;12(4):332–43.
10. Han L, Wang Y, Bryant SH. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. *BMC Bioinformatics*. 2008;9(1):401.
11. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucl Acids Res*. July 1, 2009;37(Suppl 2):W623–33.
12. Weis DC, Visco DP Jr, Faulon J-L. Data mining PubChem using a support vector machine with the Signature molecular descriptor: classification of factor XIa inhibitors. *Journal of Molecular Graphics and Modelling*. 2008;27(4):466–75.
13. Li GB, Yang LL, Feng S, et al. Discovery of novel mGluR1 antagonists: a multistep virtual screening approach based on an SVM model and a pharmacophore hypothesis significantly increases the hit rate and enrichment factor. *Bioorg Med Chem Lett*. Mar 15 2011;21(6):1736–40.
14. Franke L, Byvatov E, Werz O, Steinhilber D, Schneider P, Schneider G. Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J Med Chem*. Nov 3 2005;48(22):6997–7004.
15. Byvatov E, Schneider G. SVM-based feature selection for characterization of focused compound collections. *J Chem Inf Comput Sci*. May–Jun 2004;44(3):993–9.
16. Han LY, Ma XH, Lin HH, et al. A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *Journal of Molecular Graphics and Modelling*. 2008;26(8):1276–86.
17. Jorissen RN, Gilson MK. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J Chem Inf Model*. 2005;45(3):549–61.
18. Lepp Z, Kinoshita T, Chuman H. Screening for New Antidepressant Leads of Multiple Activities by Support Vector Machines. *Journal of Chemical Information and Modeling*. 2005;46(1):158–67.
19. Plewczynski D, Stéphane AHS, Koch U. Assessing Different Classification Methods for Virtual Screening. *Journal of Chemical Information and Modeling*. 2006;46(3):1098–106.
20. Mahmoudi N, de Julian-Ortiz J-V, Ciceron L, et al. Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening. *Journal of Antimicrobial Chemotherapy*. March 1, 2006 2006;57(3):489–97.
21. Mahmoudi N, Garcia-Domenech R, Galvez J, et al. New active drugs against liver stages of Plasmodium predicted by Molecular Topology. *Antimicrobial Agents and Chemotherapy*. January 22, 2008;AAC.01043–01007.
22. *J Chem* [computer program]. Version 5.2 Budapest: Chemaxon; 2009.
23. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Weinheim: WILEY-VCH; 2000.
24. *Dragon* [computer program]. Version 5.5. Milan, Italy: Talet; 2005.
25. Vladimir NV. *The nature of statistical learning theory*: Springer-Verlag New York, Inc.; 1995.
26. *LIBSVM: a library for support vector machines* [computer program]. Version 2.92009.
27. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)—Protein Structure*. 1975;405(2):442–51.
28. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR and Combinatorial Science*. 2007;26(5):694–701.
29. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. *Alternatives to Laboratory Animals*. 2005;33(5):445.
30. *R statistics packages* [computer program]. Version 2.8.12010.
31. Diversity Set II Information. http://dtp.nci.nih.gov/branches/dscb/div2_explanation.html. Accessed February 15, 2010.
32. Guiguemde WA, Shelat AA, Bouck D, et al. Chemical genetics of Plasmodium falciparum. *Nature*. 2010;465(7296):311–15.
33. ChEMBL—Neglected Tropical Disease. <http://www.ebi.ac.uk/chemblntd>. Accessed July 7, 2010.
34. Gamo FJ, Sanz LM, Vidal J, et al. Thousands of chemical starting points for antimalarial lead identification. *Nature*. 2010;465(7296):305–10.
35. Gagaring K, Borboa R, Francek C, et al. Genomics Institute of the Novartis Research Foundation (GNF). 10675 John Jay Hopkins Drive, San Diego CA 92121, USA and Novartis Institute for Tropical Disease, 10 Biopolis Road, Chromos # 05-01, 138 670 Singapore.



Supplementary Files

1. Supplementary1_train640.csv: Training set; compounds and molecular descriptors.
2. Supplementary2_test319.csv: Testing set; compounds and molecular descriptors.

Publish with Libertas Academica and every scientist working in your field can read your article

"I would like to say that this is the most author-friendly editing process I have experienced in over 150 publications. Thank you most sincerely."

"The communication between your staff and me has been terrific. Whenever progress is made with the manuscript, I receive notice. Quite honestly, I've never had such complete communication with a journal."

"LA is different, and hopefully represents a kind of scientific publication machinery that removes the hurdles from free flow of scientific thought."

Your paper will be:

- Available to your entire community free of charge
- Fairly and quickly peer reviewed
- Yours! You retain copyright

<http://www.la-press.com>