ISSN: 2394-2371
Review Article

A Review on Predictive Models and Classification of Inhibitors using Bioinformatics Approach

Mahima Tiwari[1], Sumit Govil[1], Shailesh Kumar[2*]

[1]Jaipur National University, Jaipur, Rajasthan, India
[2]Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur, India

ABSTRACT

Bioinformatics approach is a cost effective method for classification of Inhibitors and can be used for development of new drugs and their testing using predictive models. Here, in this review three case studies were discussed to provide insight in to the methodology used to determine a predictive model to test the drug on the basis of their molecular descriptor and how effective these models are. The inhibitors used for 5-alpha reductase 2 enzyme, Yellow Fever Virus and Epidermal Growth Factor Receptor (EGFR). Models developed for Inhibition are working very efficiently on specific class of drug. This method of modeling will be very useful for modeling various drug targets. These methods will also be helpful for drug designing which have least side effects.

Key Words- Predictive Models, Molecular Descriptor, Side effects.

INTRODUCTION

Drug development is the process to identify compounds having pharmacological properties in order to treat disease with least side effects [1]. Advancement in technology of identification and production has given need to have databases which stores the structural, biological activity and molecular descriptors of the ligands which are formulated to be used as drugs. Databases like ChEMBL and PubChem are publicly available to be used in for Virtual Screening and Docking for drug efficacy improvement. It is well recognized that the discovery of novel drugs in the pharmaceutical industry is becoming increasingly difficult, costly and time-consuming. Virtual Screening uses computational power to test large sets of chemical compounds. ChEMBL and PubChem contain both chemical structure and biological activity provides a base and opportunity to search and predict correlation for

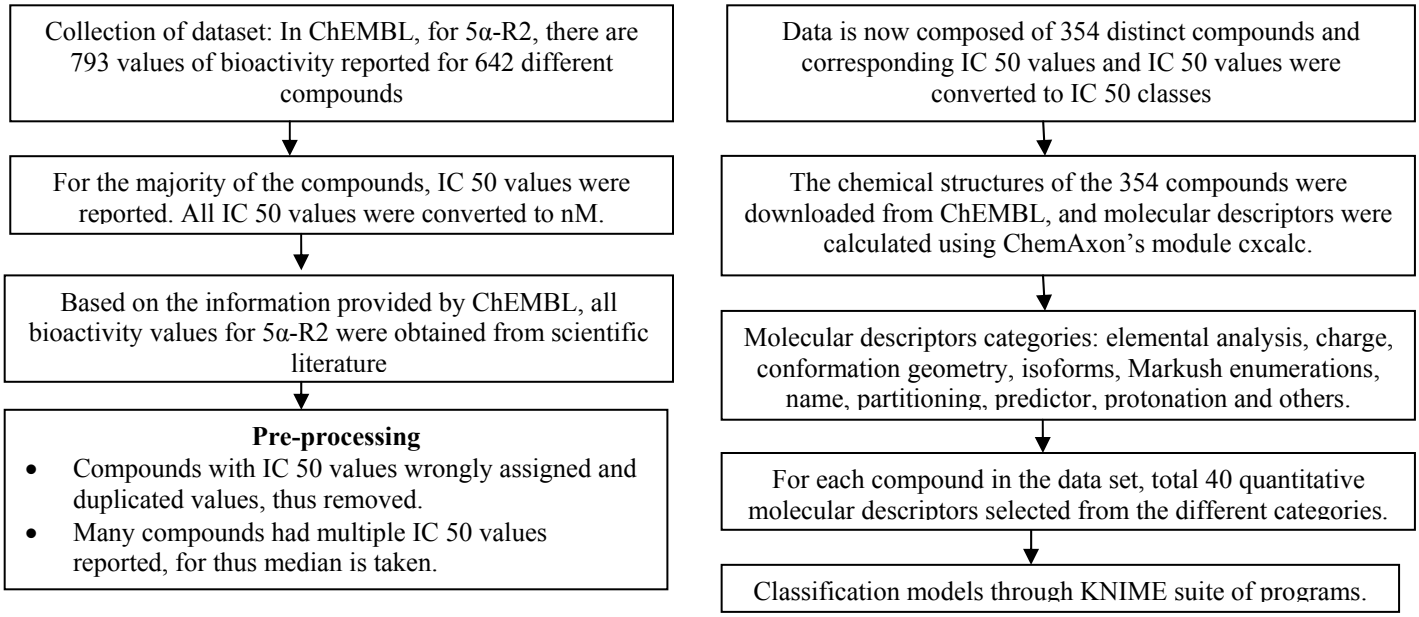
***Corresponding Author:**
Shailesh Kumar
Assistant Professor
Amity Institute of Biotechnology
Amity University Rajasthan, Jaipur, India
E.Mail: shailesh_iiita@hotmail.com
Article Received on: 05-03-2015
Revised on: 28-03-2015
Accepted/Published on: 31-03-2015

specific proteins or disease's activity in relation with activity and bio molecule structures. KNIME software [2] is used for such classification which has been developed in Java an open-source workflow platform that supports a wide range of functionality and has an active cheminformatics / bioinformatics community. In this review paper we will discuss the model prediction of three different cases for predicting the model which can distinguish between inhibitors and non-inhibitors and can enhance the process of drug development from the current available data and the models were tested for test sample.

Yellow fever [3] is spreads through mosquitoes from Yellow fever Virus which is member of Flaviviridae family, anti-YFV vaccine (17D) is commonly used in to prevent the disease. However in 5-alpha reductase enzyme [4] plays a central role in human sexual differentiation, but the increase activity of it may cause diseases such as benign prostatic hyperplasia (BPH), prostate cancer, male-pattern baldness, acne and hirsutism, thus there is a need to screen out best possible molecules which can inhibit 5-alpha reductase, the publicly available databases like ChEMBL and PubChem contain molecules which show some correlation between physico-chemical properties of bioactive molecule.

KNIME uses the data from ChEMBL and classifies the data thorugh: machine learning methods – random forests and support vector machine and Weka Data Minning Software [12] as implanted in KNIME.

Methodology Used for development of Model:



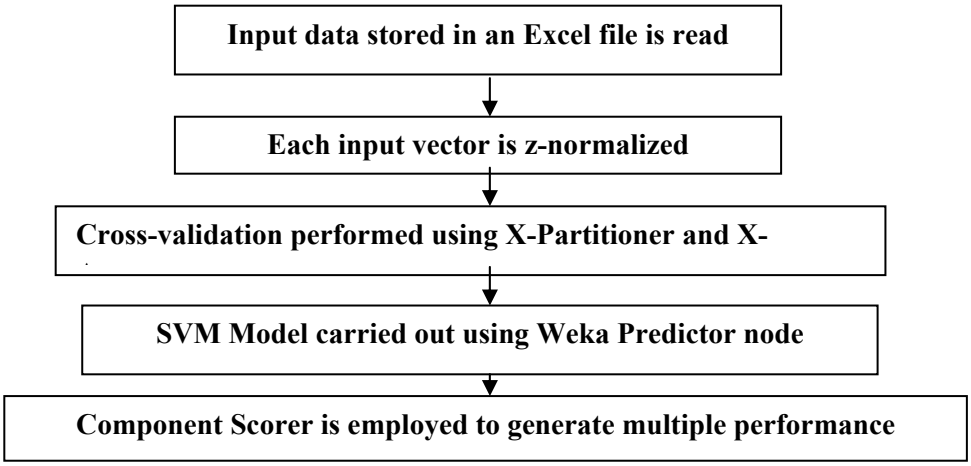
The workflow to Predictive Models for 5α-R2 inhibitors

KNIME working pipeline:

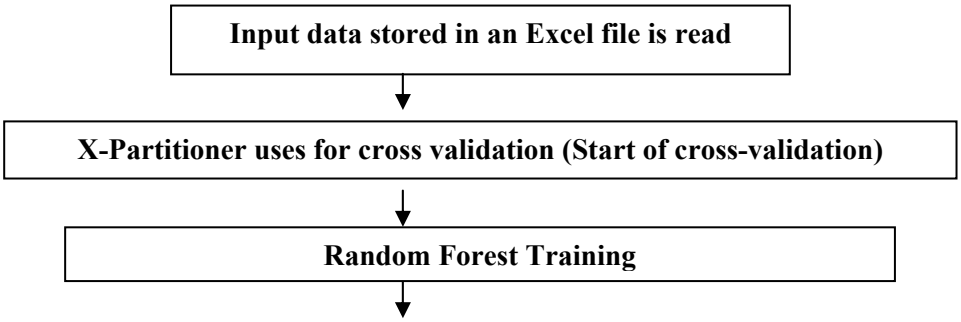
- input pre-processing
- cross-validation
- training and testing
- performance evaluation

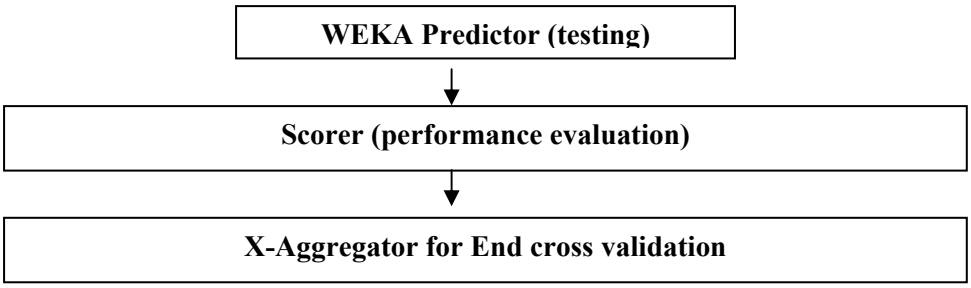
Input pre-processing, cross-validation, training and testing through two ways:

1. Support Vector Machine: SVM is a classification method which is based on the construction of a hyper plane in a multidimensional space, allowing objects in different classes to be differentiated. Data comparison through statistical method of 5-fold cross validation using random sampling.



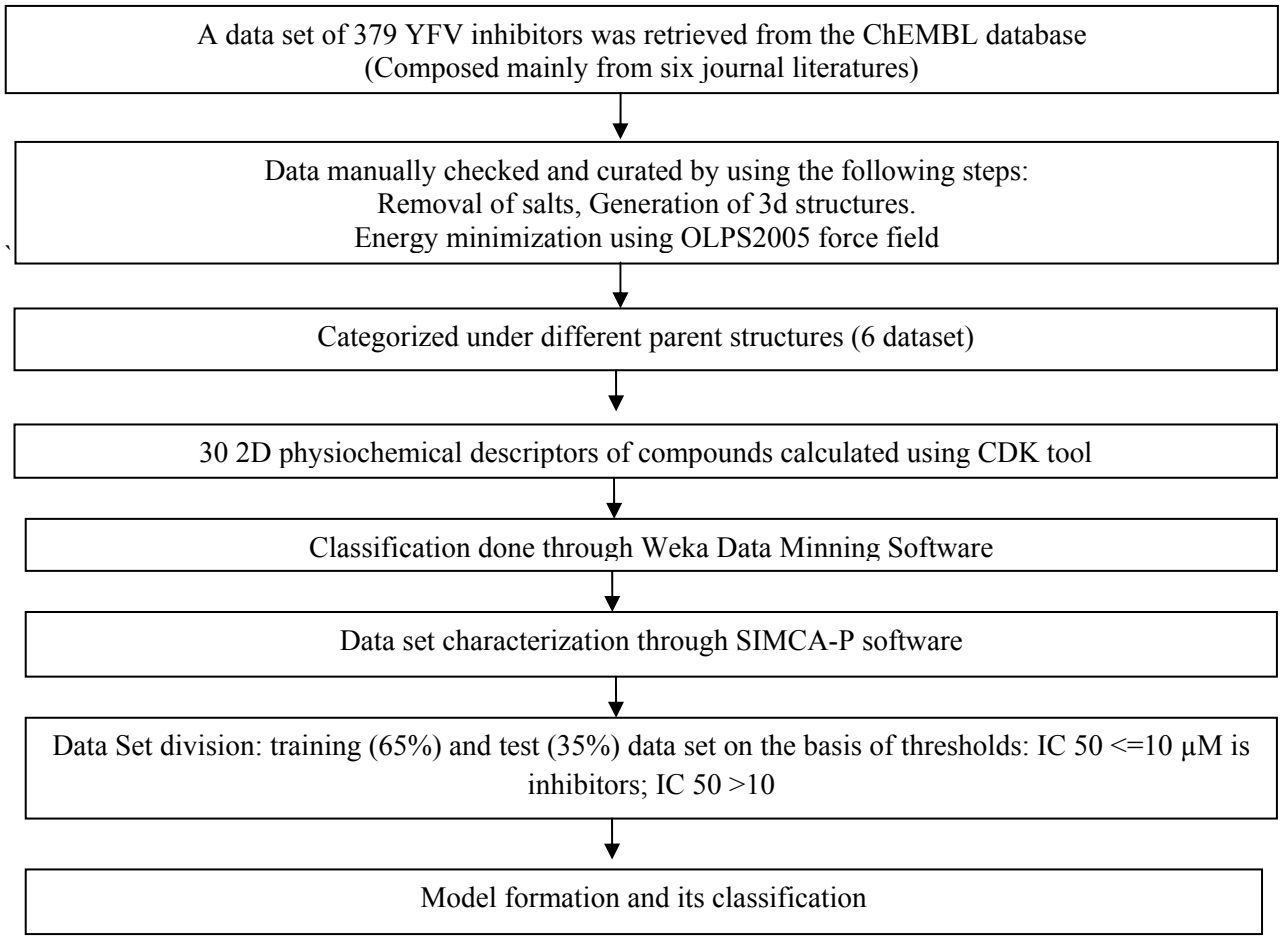
2. Random Forest: A decision tree defines a model for decisions and their possible consequences, including probabilities of outcomes, in a tree-like graph [11]. Breiman is the one who brought the concept Random Forest on the basis of decision tree. The decision tree is divided into leaves representing class labels and branches represent conjunctions of features, which is built by splitting samples based on certain variables, then repeated on each derived subsets in a recursive manner. This method is fast and gives easily interpretable results [9, 10].



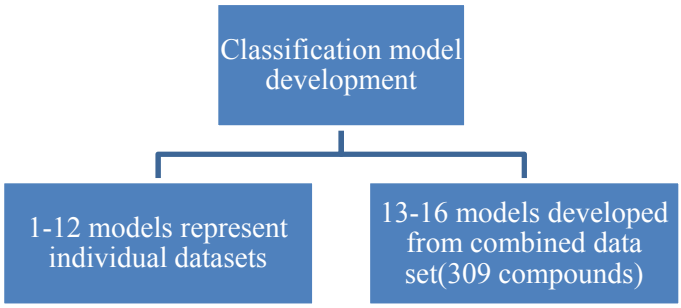


Performance Evaluation: The model generated in this work was tested in the following given parameters

Sensitivity: Sensitivity = TP/TP + FN
Specificity: Specificity =TN/TN + FP
Accuracy: Accuracy =TP + TN/TP + FP + TN + FN
Precision: Precision =TP/TP + FP
F score: F =2 × precision × sensitivity/precision + sensitivity
TP=True Positive, FN: False Negative, TN: True Negative, FP: False Positive, FN: False Negative



The workflow of KNIME Based Classification [9] Models for Yellow Fever Virus Inhibition



Model Assessment: Confusion matrix from each classification model was used to calculate various statistical parameters to assess the quality of models.

- Sensitivity=TP/TP+FN
- Specificity=TN/TN+FP
- G-mean=√Senstivity*Specificity
- MCC=(TP*TN)-(FP*FN)/$\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$
- Accuracy=TP+TN/T+N
- F-measure=Precision*Sensitivity/ Precision+Sensitivity

TP=True Positive, FN: False Negative, TN: True Negative, FP: False Positive, FN: False Negative

RESULT OF THE MODELS:

Table1: Inhibitors of 5-alpha Reductase 2

Protein	Bioactivity			# compounds	
	Total	IC_{50}	K_i	Total	Studied
5α-R2	793	466	102	642	354

Table 1. Description of the data set found for 5α-reductase (isozyme 2) in ChEMBL (accessed in December 2012) in terms of the number of bioactivity values (total, IC 50 , and K i) and number of compounds (total, and studied). [4]

Class	IC_{50} interval	# compounds
Very Good	0 – 1	107
Good	1 – 10	48
Medium	10–100	48
Bad	> 100	151

Table 2. Definition of IC 50 classes assignment.

IC_{50} Class	SVM				Random Forests			
	Sens.	Spec.	Prec.	F	Sens.	Spec.	Prec.	F
Very Good	86.9	86.1	93.9	86.5	85	85	93.5	85
Good	56.2	55.1	92.8	55.7	60.4	54.7	92.2	57.4
Medium	14.6	46.7	97.4	22.2	31.2	50	95.1	38.5
Bad	90.1	74.7	77.3	81.7	88.1	88.1	84.7	84.4
Accuracy	74.3				75.7			

Table 3. Evaluation of the classification models for 5α-R2 using 5-fold cross validation. The different performance measurements: sensitivity (Sens.), specificity (Spec.), precision (Prec.), F score (F) and accuracy are shown for SVM and random forests learning algorithms. All values are shown in percentage (%). [4]

Yellow Fever Virus Inhibition:

Characterization of Dataset: Principle Component Analysis was done to check possible clusters, outliers, similarities or dissimilarities, distribution of inhibitors and non-inhibitors in the training and test set in the physicochemical space. The result showed that the diversity of dataset is satisfactorily reflected in the training set and there are no distinct clusters in the dataset. There are some distinct outliers were observed. The result also showed that inhibitors are highly influenced by topological polar surface area (TPSA) and polar bonds. This reveals that non-inhibitors are relatively more hydrophobic than inhibitors [8]. The classified models according to Naïve Bayes [5] showed overall accuracy test set is >75%. Except the two dataset (-0.11) the rest showed 0.6 value in terms of MCC for quality in model. This poor performance was due to the fact that there were only two non-inhibitors, which were predicted as inhibitors because of not only sharing of common scaffolds as positive but also shares similar structural patterns.

In order to provide the YFV inhibition model to the medicinal chemistry community, model 15 is implemented into KNIME workflow

Inhibitors of 5-alpha reductase 2 enzyme:

The only two currently marketed inhibitors (Finasteride and Dutasteride) cause undesirable side effects, stressing the need to search for more potent and selective inhibitors. The public datasets help in in-silico screening and thus prediction of more potent inhibitor of 5-alpha Reductase 2, in which both SVM and random forests can be used to distinguish between compounds with very good and bad IC 50 values.

In this study, the KNIME based classification models were developed using the existing YFV inhibitors from the ChEMBL database [4]. The best classification model is able to discriminate >90% of inhibitors from non-inhibitors with an overall accuracy of >90%. Subsequently, the best model is implemented in the KNIME workflow which could be used as a virtual screening workflow to screen novel molecules for the YFV inhibitory activity.

Inhibitors of Epidermal Growth Factor Receptor (EGFR) enzyme:

Epidermal Growth Factor Receptor (EGFR) is a most studied cancer drug target. The treatment of patients with EGFR based inhibitors as targeted therapy thus has shown a significant reduction in the cancer progression Various QSAR models have been developed for predicting inhibition activity of molecules against this receptor. Previous models were for only limited set of molecules of class like

quinazoline-derivatives. In this study, an attempt has been made to develop prediction models on a large set of molecules (~3500 molecules) that include diverse scaffolds like quinazoline, pyrimidine, quinoline and indole [6]. Random forest based model achieved maximum Matthew's correlation coefficient (MCC) 0.49 with accuracy 83.7% on a validation set using 881 PubChem fingerprints. This is an important study for development of new cancer inhibitors.

REFERENCES

1. Ewa Bielska, Xavier Lucas, Anna Czerwoniec, Joanna M. Kasprzak, Katarzyna H. Kaminska, Janusz M. Bujnicki. Journal of Biotechnology, Computational Biology and Bionanotechnology. Review Paper. 2011. vol. 92(3). pp. 249-264

2. Stephan Beisken, Thorsten Meinl, Bernd Wiswedel, Luis F de Figueiredo, Michael Berthold and Christoph Steinbeck. BMC Bioinformatics. 2013. Vol.(14:257).pp. 1471-2105.

3. N.S. H. N.Moorthy and V. Poongavanam, RSC Adv., 2015, DOI: 10.1039/C4RA15317K.

4. In Search of Predictive Models for Inhibitors of 5-alpha Reductase 2 Based on the Integration of Bioactivity and Molecular Descriptors Data. Joana Sousa, Rui M. M. Brito, Jorge A. R. Salvador and Cândida G. Silva. Proceedings IWBBIO 2014. 2014.pp.464-472

5. S. Beisken, T. Meinl, B. Wiswedel, L. F. de Figueiredo, M. Berthold, C. Steinbeck, BMC Bioinformatics, 2013, 14, 257.

6. Harinder Singh, Sandeep Singh, Deepak Singla, Subhash M Agarwal and Gajendra P S Raghava, QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest, Biology Direct (2015) 10:10

7. T. J. Chambers, A. D. Droll, Y. Tang, Y. Liang, V. K. Ganesh, K. H. M. Murthy, M.Nickells, J. Gen. Virol., 2005, 86, 1403–1413.

8. N. S. H. N. Moorthy, N. M. F. S. A. Cerquira, M. J. Ramos, P. A. Fernandes, Chemom. Intellegent Lab. Sys., 2015, 140, 102-116.

9. K. K. Chohan, S. W. Paine, J. Mistry, P. Barton, A. M. Davis, J. Med. Chem., 2005,48, 5154–5161.

10. N. S. H. N. Moorthy, S. F. Sousa, M. J. Ramos, P. A. Fernandes, RSC Adv., 2014, 4(106), 61624–61630.

11. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

12. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using WEKA. Bioinformatics. 2004;20:2479–81.