ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 7, Issue, 4, pp. 10370-10373, April, 2016

International Journal of Recent Scientific Research

Research Article

PREDICTIVE MODEL ASSISTED IN SILICO SCREENING OF ANTI-LUNG CANCER ACTIVITY OF COMPOUNDS FROM LICHEN SOURCE

Mahesha Nand[1]., Priyanka Maiti [2]., Ragini Pant [3]., Subhash Chandra* and Veena Pande[4]

[1,3,4]Department of Biotechnology, Kumaun University, Bhimtal Campus Bhimtal, Uttarakhand, India
*,[2]Department of Botany, Kumaun University, S.S.J Campus, Almora, Uttarakhand, India

ARTICLE INFO

Article History:

Received 16th January, 2016
Received in revised form 24th February, 2016
Accepted 23rd March, 2016
Published online 28th April, 2016

Keywords:

Machine Learning, Data Mining, WEKA, Classification, Lichen, Docking, Pharmacophores.

ABSTRACT

Lichen derived compounds have been reported to have various therapeutic potentials. The present work explores anti-cancer potential of such compounds against human Non-Small Cell Lung Cancer cell line NCI-H322M by using computational methods. Initially a predictive model was developed based on machine learning approach using several classifiers like Random Forest, J48, Bagging, PART and Random Tree in WEKA software. Random tree classifier showed its potency in terms of sensitivity (1), specificity (1), time (.02 sec) and other parameters. This model screened nineteen compounds with active potential out of seventy lichen compounds. Docking simulations were further performed to evaluate the binding potential of those molecules with the tyrosine kinase domain of epidermal growth factor receptor using AutoDock Vina. Four compounds namely Asperphenamate (-9.9 kcal/mol), Brefeldin A(-9.2 kcal/mol), Simvastatin (-10.2 kcal/mol) and Gliotoxin (-8.4 kcal/mol) showed excellent binding potential in the Erlotinib (-8.2 kcal/mol) binding site where as five compounds namely Aculeatins A (-5.6 kcal/mol), Brefeldin A (-6.6 kcal/mol), Compactin (-6.3 kcal/mol), Wortmannin (-9.7 kcal/mol),Phaeosphaerin B (-7.7 kcal/mol)showed their potential in Gefitinib (-7.9 kcal/mol) binding site. Additionally pharmacophores evaluation was done on the screened molecules to compare them with common pharmacophores of Erlotinib and Gafitinib. Results reflect presence of considerable number of heterocycle fragments in the screened compounds with accepted range of other pharmacological properties.

INTRODUCTION

Lung cancer is the leading cause of cancer death worldwide, with an estimated 221,200 new diagnoses and 158,040 deaths in 2015, according to the FDA. The WHO reports reflects that among all types of cancers occurrence of lung cancer is most common and responsible for 1.59 million deaths in 2012 (WHO, 2012). Eastern Europe has the maximum lung cancer mortality among men, while northern Europe and the U.S. have the uppermost mortality among women (SEER, 2010).

At present time high throughput screening (HTS) is widely used to screen compounds based on chemical structure features and other diverse biological features of known set of compounds (Teixeira *et al.*, 2013). The chief objectives of machine learning focus on the exploration of huge data sets with non–congeneric molecules. The logic behind the custom of machine learning is to determine patterns and signatures in data sets from high through put in-vitro assays (Witten *et al.*, 2011). The Weka workbench is an assembly of contemporary machine learning algorithms with data preprocessing tools. Weka stands for Waikato Environment for Knowledge Analysis and was originally developed at the University of Waikato in New Zealand. The system is scripted in Java and dispersed under the terms of the GNU General Public License. Weka extensively runs on nearly any platform and has been tested under Linux, Windows, and Macintosh operating systems. It offers a harmonized interface to many diverse learning algorithms, accompanied by methods for pre and post processing and for assessing the result of learning schemes on any specified dataset (Gewehr *et al.*, 2007).

Present study deals with the development of machine learning model in Weka to screen compounds against human Non-Small Cell Lung Cancer cell line NCI-H322M. In vitro experiment is crucial for screening of compounds active against specific target/cell line. Generally, different human cancer cell lines are utilized in *in vitro* experiment for the screening of anticancer compounds (Liao *et al.*, 2012). Likewise, in virtual screening, screening of compounds is performed by building models

*Corresponding author: **Subhash Chandra***

based on previously known data set to predict the possible activity of compounds. In the present study, supervised machine learning based model is used for *in silico* assay. The supervision to the model has been supplied in the form of percentage inhibition on human Non-Small Cell Lung Cancer cell line NCI-H322M. The model was applied to screen a library of seventy compounds obtained from lichens. Lichens are slow growing symbiotic organisms which produce a diverse and wide range of secondary metabolites with diverse biological activities (Johnson *et al.*, 2011). The heterocyclic compounds and common heterocycle fragments present in most pharmaceuticals presently marketed are well reported for their intrinsic versatility and exclusive physicochemical properties. One of the rich store house of these compounds are Lichens. The biological potential of many compounds from lichens has largely remained unexplored. So the aim of the study deals with the unknown /hidden compounds from lichens which have the potential to inhibit lung cancer.

MATERIALS AND METHODS

Data set retrieval and pre- processing

In present study, the complete data set of percentage inhibition for Non-Small Cell Lung Cancer cell line NCI-H322M, was retrieved from CHEMBL database (Gaulton *et al.*, 2011). The assay (CHEMBL3301526) consisted of a total of 400 compounds capable of inhibiting Lung Cancer cell line NCI-H322M cell line. All compounds were categorized based on inhibition percentage score. The compounds that had an activity score between 90% and 127.63% were defined as active and the compounds having an activity score between 89.9 % and -94.4 % were defined as inactive. The Structural Data Format (SDF) of the molecules was obtained from their simplified molecular-input line-entry system (SMILES) format using Open Babel software (Liew *et al.*, 2012).

Chemical descriptor calculation

Descriptors for the above dataset were calculated using PaDEL- software (Hall *et al*, 2009).The software calculates 797 descriptors (663 1D, 2D descriptors, and 134 3D descriptors) and 1 type (PubchemFingerprinter) of fingerprint, using the Chemistry Development Kit. The bit-string fingerprint attributes of only all 0's value (all 0's or all 1's) all across the molecules were removed to reduce the dimensionality of the dataset. The complete set of molecules was randomly divided into 20% independent test set and 80% training validation set.

Machine learning classifiers

In present work, Weka (Dubey *et al.*, 2011) open source software was used to build machine learning models. Different classifiers are available in Weka for data mining and machine learning. It has various tools for data pre-processing, classification, regression, clustering, association rules, and visualization. In present study predictive models were built using five classifiers namely Random Forest, J48, Bagging, PART and Random Tree respectively.

Random forest can defined as sequence of tree predictors where every single tree varies on the values of a random vector sampled individually and with the identical distribution for each trees in the forest. These classifiers combine "bagging" concept and the random selection of features, in order to build collection of decision trees with measured variation. By using bagging mechanism RF builds an ensemble of CART tree classifications. Each node of trees merely chooses a small subset of features for the split and empowers the algorithm to generate classifiers for high dimensional data very rapidly. RF runs proficiently on huge data sets with a lot of features and its execution speed is fast (Witten *et al.*, 2011).

J48 algorithm is mainly based on J.R. Quilan C4.5 algorithm. Only categorical type data are examined through this algorithm. The algorithm built a C4.5 decision tree and each time when it executes, an instance of the class is created by allocating memory for constructing and storing the decision tree classifier. The J48 class does not accurately include any code for constructing a decision tree. It incorporates references to instances of other classes that do mainly the work. The decision tree is made by gaining the information from training data and the attribute with highest normalized information (Breiman, L., 1996).

Bootstrap Aggregating or **Bagging** is one of the most effective ensembles learning method in machine learning and was developed by Breiman. It generates different classifiers and then ensemble them to select certain base classifier algorithm. After that these base classifiers are trained on random redistribution training datasets (Hall *et al.*, 2009).

PART is a One Class Classifier. This classifier decreases the class of existing classified data to just a single class, and pick up the data using any information from other classes (Pfahringer B., 2010).

Random Tree can be categorized as supervised Classifier. This classifier is a combination of two preexisting algorithms, single model trees and Random Forest. It utilizes a bagging idea to create a random set of data for constructing a decision tree. In standard tree each node is divided using the best split among all variables. The beauty of this algorithm is that it can deal with both classification and regression problems. Initially the classifier shots the input feature vector and classifies it with each tree in the forest. Finally it outputs the class label that obtained the majority of "votes".

Revalidation of model and screening of compounds

All the nineteen molecules with active potential were docked with EGFR protein in two different binding sites of two different drugs (Erlotinib and Gefitinib) to check their ability against the protein. Docking studies were carried out using AutoDock in PyRx. PyRx (Trott O. Olson A J, 2010) is a graphical user interface for AutoDock 4.2 and AutoDock Vina for performing virtual screening. Molecular Docking was performed to calculate the favored orientation of ligand molecules with EGFR when bound with each other to form a stable complex. AutoDock 4 contains of two core programs: AutoGrid for calculating the grids and AutoDock for performing docking of a ligand to a set of grids describing the target protein. The three dimensional structures of the protein were obtained from protein data bank. The pdb ID for Erlotinib binding site was 4HJO and Gefitinib binding site was 4WKQ. The proteins and ligand molecules ware prepared, optimized and grid is generated near the active site of protein molecule

using AutoDock tools and then docking was performed. Subsequently the docked poses were analyzed using PyMOL software. The output showed binding affinities and RMSD scores for each ligand with nine different poses.

RESULTS AND DISCUSSION

Using Weka a number of models were generated by applying distinct classifiers to examine the performance of the selected classification methods or algorithms namely Random Forest, J48, Bagging, PART, Random Tree. The best model was selected on the basis of several statistical parameters. All the models had accuracy greater than 90%. The performance of the models were assessed by evaluating their several sub parameters like Correctly Classified Instances % (value), Incorrectly Classified Instances % (Value), Time Taken (seconds) and Kappa Statistic (Table-1).

Table 1 Different statistical parameters for models made by different classifiers

Algorithm	Random Forest	J48	Bagging	PART	Random Tree
Correctly Classified Instances %(Value)	98.98	97.72	91.64	97.97	100
Incorrectly Classified Instances % (Value)	1.01	2.27	8.35	2.02	0
Time Taken (seconds)	0.03	0.04	0.02	0.02	0.02
Kappa Statistic	0.97	0.95	0.82	0.95	1

The results of the table clearly indicate that Random tree was the best one in terms of accuracy and time. Several other statistical measures such as sensitivity and specificity were considered to judge the robustness of the models. The outcomes indicated that all the models had sensitivity and specificity greater than 90% but the Random tree model was best due to having highest sensitivity value and Bagging model had the least (Figure -1). The best model was evaluated on 6 PubChem bioassays to check its efficiency. Four selected bioassays were on Growth inhibition of NCI-H322M cells after 48 hrs. by sulforhodamine B assay (AID:

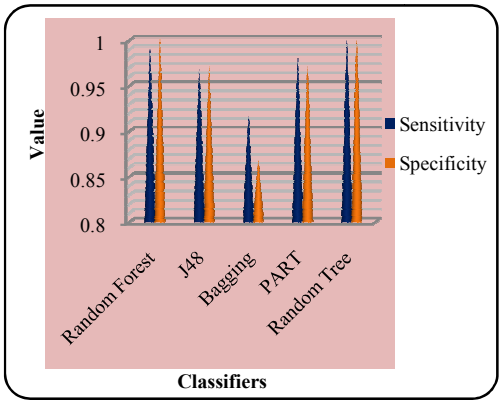

Figure- 1 Sensitivity and Specificity measurement of different classifiers

670653, AID: 538167, AID: 352337, AID: 538287) and two respective bioassays were on Growth inhibition of human NCI-H322M cells (AID: 1125045) and *In vitro* total growth inhibition against non-small-cell lung cancer NCI-H322M cell lines (AID: 147580). Figure 2 is showing the comparison between the functional capabilities of the model for sorting the active compounds with the original active compounds from the bioassay. The efficiency of the model was clearly reflected from the results. Ultimately the selected model was used to screen a library of 70 compounds derived from Lichens. Nineteen compounds (Figure -3) were obtained with active potential against non-small-cell lung cancer NCI-H322M cells.

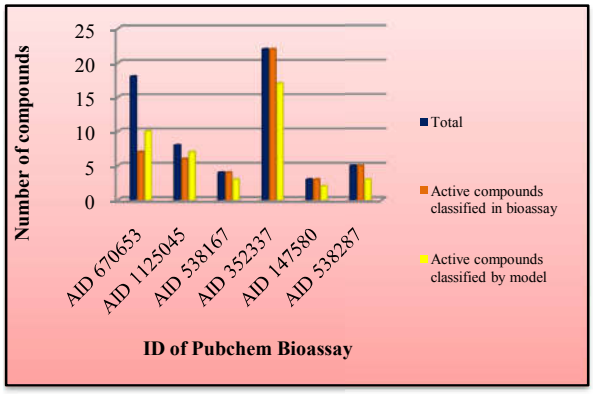

Figure 2-Validation of the model in PubChem Bioassays

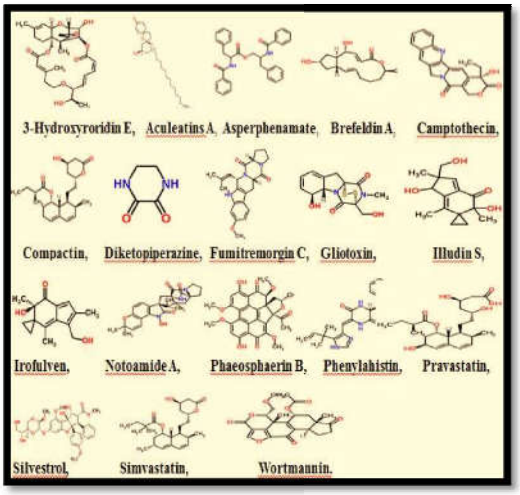

Figure 3 -Structures of screened compounds

Tyrosine kinase domain of EGFR protein with two binding sites for Erlotinib and Gefitinib were analyzed in AutoDock vina software. The binding site search showed that Gefitinib binding involves the amino acids LEU 718 A - THR 854 A with 55 non-bonded contacts.

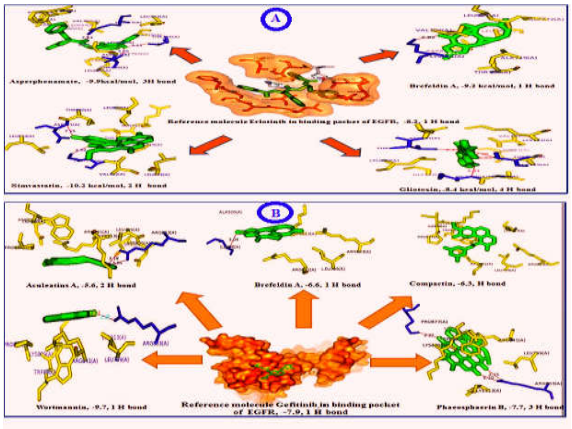

Figure-4 Binding conformation of Hit molecules in A: Erlotinib, B: Gefitinib binding site

In case of Erlotinib the interacting residues are LEU 694 A-ASP 831 A involving 64 non-bonded contacts and 1 hydrogen bond. All screened compounds were docked in both binding pockets and the results revealed that four compounds namely Asperphenamate, Brefeldin A, Simvastatin and Gliotoxin showed their efficiency with excellent binding energy scores respectively -9.9 kcal/mol, -9.2 kcal/mol, -8.4 kcal/mol and -10.2 kcal/mol when compared with Erlotinib (-8.2 kcal/mol) (Figure – 4,A). In case of Gefitinib binding site, five compounds Aculeatins A, Brefeldin A, Compactin, Wortmannin and Phaeosphaerin B showed acceptable range of binding energy -7.7 kcal/ mol, -5.6kcal/mol, -9.7 kcal/mol, -6.6kcal/mol and -6.3 kcal/mol respectively (Figure – 4, B).

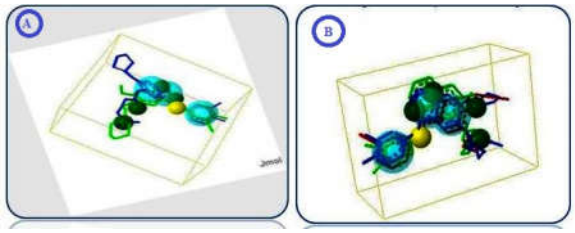

Figure -5 Common Pharmacophores A: Erlotinib, Gefitinib, Afanitib ,B: Hit molecules

Hydrogen bond acceptor
Hydrophobic group
Aromatic rings

Finally the molecules were analyzed for their common pharmacophores by pharmaGist server. Detection of pharmacophores were done by extracting common chemical features from 3D structures of the active ligand set (Erlotinib, Gefitinib, Afanitib) that are representative of essential interactions between the ligands and EGFR. Certain elucidated physico-chemical properties were hydrogen-bond acceptor/donor atom; set of aromatic rings and adjacent hydrophobic atoms. The common pharmacophores from Erlotinib, Gefitinib, Afanitib were 3 hydrogen bond acceptors, one hydrogen bond donor, three aromatic rings whereas the common pharmacophores from screened hit molecules showed three hydrogen bond acceptors, zero hydrogen bond donor and three aromatic rings. So the results clearly reflect the presence of heterocyclic rings in the screened molecules which increases their significance in terms of drug likeness as nearly 60% of unique small-molecule drugs approved by FDA contains heterocyclic rings.

Acknowledgements

The authors are thankful to The Head, Department of Botany, Kumaun University, SSJ Campus, Almora, for providing necessary facilities to carry out the present study.

References

Breiman, L., 1996. Bagging predictors, Machine Learning, 24:123-140.

Dubey, A., Pant, B., Chouhan, U, 2011. Machine learning model for HIV1 and HIV2 enzyme secondary structure classification, J. Comput. Method. Mol. Design, 1 (2): 1-8.

Gaulton, A; *et al*. 2011. "ChEMBL: a large-scale bioactivity database for drug discovery". Nucleic Acids Research 40: D1100–7. doi:10.1093/nar/gkr777. PMC 3245175. PMID 21948594O'Boyle, N.M., Banck, M, James, C.A, *et al*., 2011. Open Babel: An open chemical tool box, Journal of Chemo informatics, 3: 33.

Gewehr, J.E., Szugat, M, Zimmer, R.2007.BioWeka Extending the Weka Framework for Bioinformatics, Bioinformatics, 23 (5): 651-653.

Hall, M., Frank, E, Holmes, G, Pfahringer, B, Reutemann, P, *et al*., 2009.The WEKA data mining software. ACM SIGKDD Explore News, 11: 10.

Johnson, C.J., Bennett, J.P, Biro, S.M, Duque-Velasquez, J.C, Rodriguez, C.M, Bessen, R.A, *et al*., 2011. Degradation of the disease-associated prion protein by a serine protease from lichens, PLOS One, 6:19836.

Liao, W.Y., Shen, C.N, Lin, L.H, Yang, Y.L, Han, H.Y, *et al*., 2012. Asperjinone, a nor-neolignan, and terrein, a suppressor of ABCG2-expressing breast cancer cells, from thermophilic Aspergillus terreus, J Nat Prod, 75: 630–635.

Liew, H.Y., Sharma, C.Y.N, Woo, S.K, Chau, Y.T, Yap, C.W, 2012.PaDEL DD Predictor Open-source software for PD-PK-T prediction, *Journal of Computer Chemistry*, 4:1-15.

National Cancer Institute, SEER stat fact sheets, 2010. Lung and Bronchus. Surveillance Epidemiology and End Results.

Pfahringer B., 2010. Random model trees, an effective and scalable regression method, The University of Waikato, Computing and Mathematical Sciences, New Zealand: 1:9.

Teixeira, A.L., Leal, J.P, Falcao, A.O.2013.Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons, *Journal of Cheminformatics*, 5 (9): 2-15.

Trott O. Olson A J, 2010, AutoDock Vina Improving the speed and accuracy of docking with a new scoring function efficient optimization and multithreading, J Computed Chem. 31:455–461.

Witten, I.H., Frank, E, Hall, M.A.2011.Data Mining Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, pp-1-10.

World health organization (WHO). Fact of the Cancer.2012. Available at http://who.int/mediacentre/factsheets/fs297/en/index.html. Accessed 4/3/2016.

****

How to cite this article:

Mahesha Nand *et al*.2016, Predictive Model Assisted In Silico Screening of Anti-Lung Cancer Activity of Compounds from Lichen Source. *Int J Recent Sci Res.* 7(4), pp. 10370-10373.