Analysis Of The analog Drugs For The Cancer Inhibitor Alisertib

Prajna Manoor Kumar
prajnamk05.pmkmk@gmail.com
University of Central Florida, Orlando, Florida

Sunny Kotwal
sunnykotwal1234@gmail.com
University of Central Florida, Orlando, Florida

ABSTRACT

Alisertib is a type of inhibitor that suppresses the expression of the Aurora A kinase enzyme , whose over-expression leads to cancer in human body. An extensive literature review was carried out to understand the risks associated with the inhibitor and also the alternative solutions that the researchers have came up with. At present analogs of this inhibitor is being used in treatment of cancer. Over the years there are a number of analogs that have been introduced to market. These analogs may or may be similar to original inhibitor.
At present there is no readily available solution for analyzing the similarities of these drugs. Here is a paper that would be helpful in testing the efficiency of the analog drugs. In this paper a comparative study of performance of the machine learning algorithm to chose the best fit algorithm for the drug analysis has been done, that would be helpful in developing a machine learning model in the near future.

KEYWORDS

Dataset, kinase, inhibitor, analog, supervised, unsupervised,enzyme,lymphoma

1 INTRODUCTION

Kinase is a protein that controls most of the signals in cells including apoptosis, the cell cycle and differentiation.[1]Mitosis has important functions controlled by three types of Aurora Kinases(AK)[2] The mitotic processes that control kinetochores and spindles are from type A of Aurora kinase(AAK).[3] When levels are too high, AAK causes cancer transformations and mutations in the chromosomes.[4] This overexpression has been seen to form multiple types of cancers of the head, lungs and breast.[5,6] AAK overexpression is also linked to aggressive non-Hodgkin lymphomas, with very high rates of risks of both B-cell and Burkitt lymphoma are seen.[7]

A drug that was being developed, Alisertib, is a strong inhibitor of AAK that could slow down mitosis and was shown to reduce cancer expansion.[8] In early phase studies after passing for safety it was being used for pediatric and non-hematologic cancers.[9] Tests in conjunction with various chemotherapies were found to be handled well.[10] Specific chemotherapies such as vincristine and rituximab with Alsertib had massive cell death in B-cell lymphoma.[11] Unfortunately when it hit phase three the results were no longer promising with too many risks associated with the drug and research into other drugs was prioritized.

For machine learning to make accurate prediction it is imperative to have reliable data sets that are far larger than what could be perceived a short few decades ago.[12] Cancer data sets are doubling every 6 months by just genomic studies and in less than a decade will reach up to forty million gigabytes each year.[13] The data set that will be used is with Alisertib analogs. An Analog is a compound that has a similar structure of another but has minor differences in its form or molecular framework. [14] Several machine learning algorithms will be performed on the data set to see what correlations can be found between the analogs.

2 METHODOLOGY

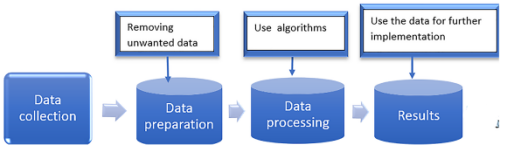

Figure 1: method architecture

2.1 Data Collection

For this project, the data on various analogs of Alisertib inhibitor were needed in order to predict the similarity of the compounds. This data can be collected from and database system called PubChem, a database system for chemical molecules. It contains information regarding various chemicals and their activities against the biological samples. This database system is

maintained by the National Center for Biotechnology Information(NCBI), a part of National Library of Medicine, which intern is a part of United States National Institutes of Health. This database system is a open database system which means we can input the details of the compound and later anyone can use. In PubChem webpage on entering the compound we will get the original compound, on keeping the similarity rate to 70%, we have obtained more than 1000 analog for the compound Alisertib. These analog vary in terms of the their composition and molecular weight.

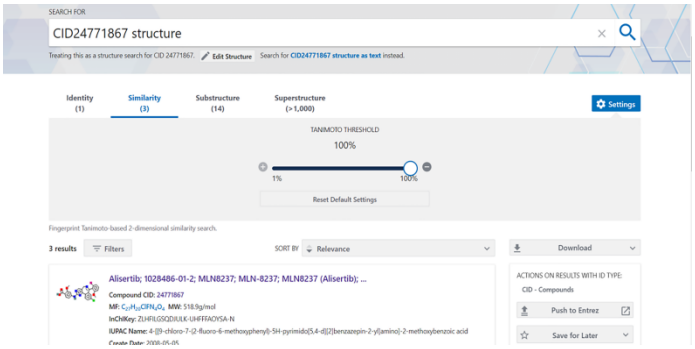

Figure 2: PubChem database system

2.2 Data Preparation

In this model we need to find the similarity of the analogue compounds. Before the data is fed into to the model, it has to be prepared and for data preparation we have used spyder, which is a free python integrated developmental environment. This IDE is included with Anaconda. This platform allows the editing debugging and even testing of the data.

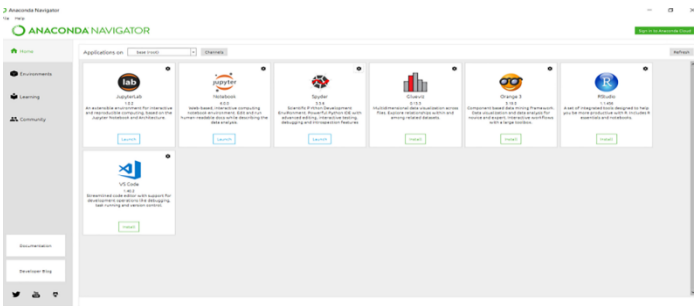

Figure 3: Jupyter user interface

In the process of data preparation the, cleansing of the dataset is done. It involves the removal of the unwanted data, unimportant columns and also correction of the missing values.

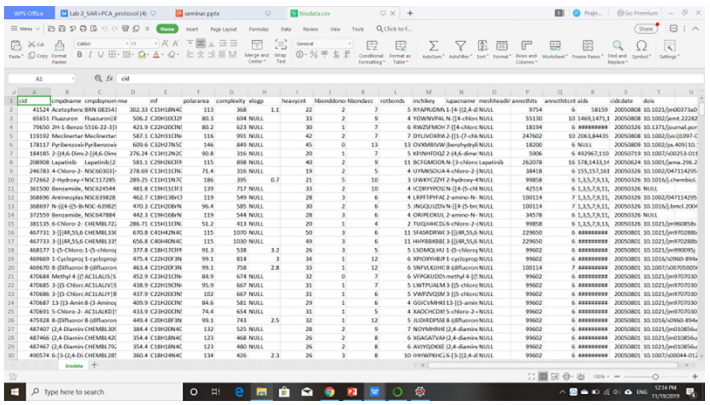

Figure 4:Dataset before cleansing

Above figure shows the actual dataset. The entire dataset is loaded into Spyder. The dataset has many unreliable columns. Cmpdsynonym, inchikey, iupacname, meshheadings, aids, cidcdate, dois,cmpdname, mf, cid, xlogp are the set of columns that had to dropped before the data processing was applied. The original data consisted of 20 columns and 1002 rows and after processing it is reduced to 1000 rows and 9 columns and these values are converted to floating point. The image below shows the prepared dataset.

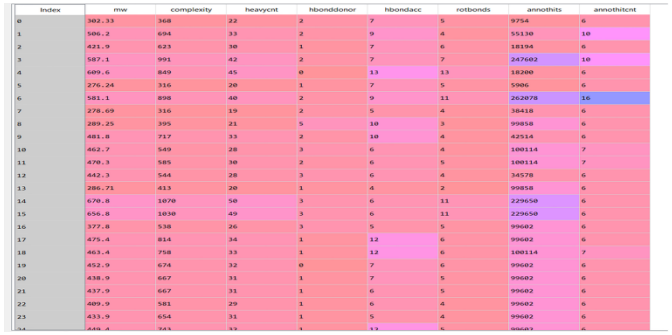

Figure 5: Dataset after pre-processing

2.3 Data Processing

At this stage we implement the machine learning algorithm to our dataset to test which algorithm suits the best. This process involves the application supervised and unsupervised set of algorithms. We started of with supervised algorithms, where we considered molecular weight as the dependent feature based on which the analysis was done. The algorithms used are listed below.

- Logistic Regression:

It is statistical model used and to determine the relation between the one more dependent variable to the independent variable. It considers the dependent variable with two possible values like pass/fail, these values are called as log-odds and the model converts these log-odds to probability.

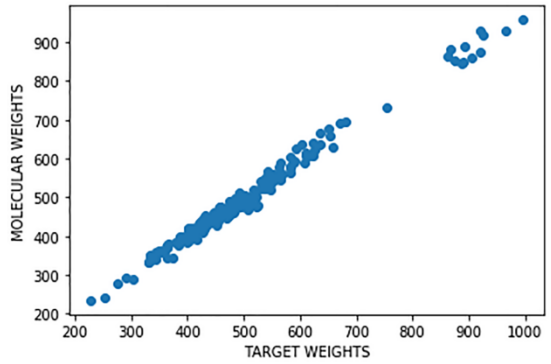

Figure 6: Scatter plot forLogistic Regression

When applied on the analogue dataset, above graph was obtained with 98.67% accuracy.

- SVM:

It is a type of supervised learning in which the categorical classification is done. It determines a hyper-plane that clearly determines the set of values belongs which class. It allows transformations called as kernels, they are of 4 types that is linear, rbf, polynomial, sigmoid. In this project we have used linear kernel.

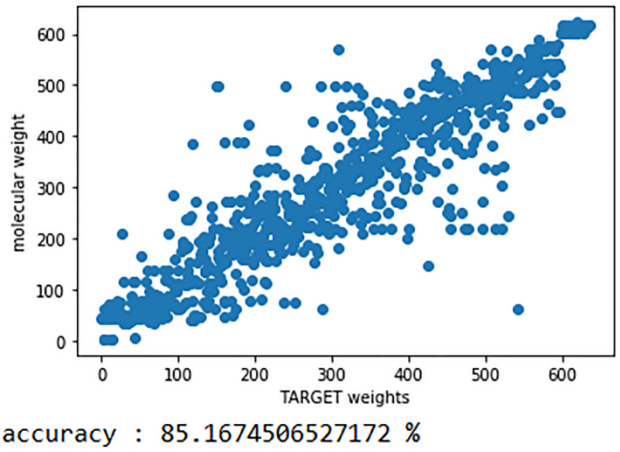

Figure 7: Scatter plot of SVM

The above figure shows the graph obtained on implementing the algorithm on analogue dataset.

- KNN:

It is one of the simplest classification algorithm. It is also called as lazy learning algorithm. It is good in predicting the classification for the database in which the values are separated into several classes. This algorithm does not depend on the under lying data distribution, hence this algorithm is much useful in real world application.

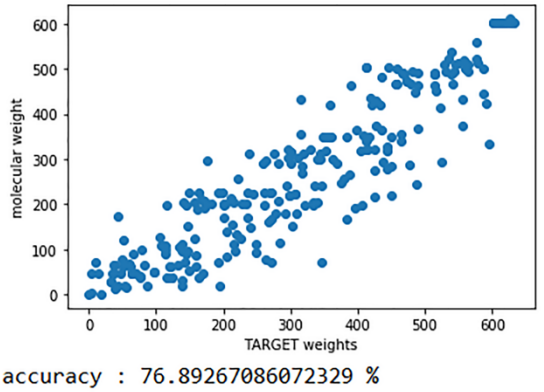

Figure 8: Scatter plot of KNN

Above figure shows the plot of the output for the analogue drug data.These are the set of the algorithms used in supervised learning. They we jumped into unsupervised learning.Clustering is technique of grouping all the similar values into a group in such a way that the data values in one group is are more similar to their neighbors in the same group than the ones in the another group. These data groups are called as clusters.

- Hierarchical clustering:

Hierarchical clustering is method of clustering analysis in which hierarchy of cluster is formed. This method falls into two types,
Agglomerative clustering:
Its is bottom up approach where all the observation starts in it's own cluster and then the clusters are merged as it moves up the hierarchy.
Divisive clustering:
It is top down approach, here all the observations starts with one cluster and then its moves down hierarchy to give out the final cluster values.
In this project we have applied the agglomerative clustering and the output of this method is shown in the below diagram.

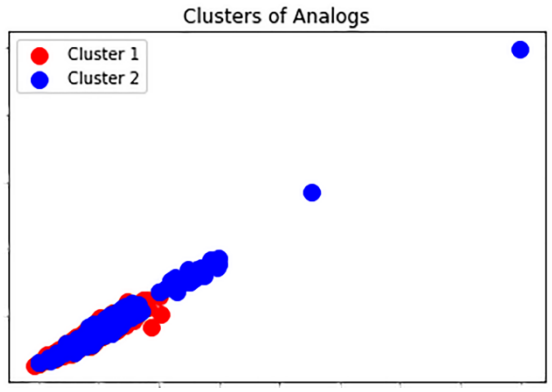

Figure 9: Output of hierarchical clustering

- Gaussian Mixture Model:

A probabilistic model for representing normally distributed set of values within the overall population. The model does not require

to have a prior knowledge on, which group a data point is originally belongs to. The model gains the knowledge by the maximum likelihood estimation technique which seek to maximize the probability. Below image shows the analysis result for the drug dataset.

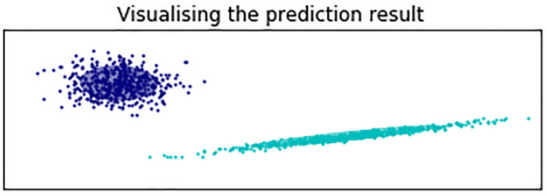

Figure 10: Output of GMM

3 RESULTS AND DISCUSSIONS:

Table 1 : Accuracy of supervised learning algorithms

Algorithm	Accuracy
Logistic Regression	98.67%
SVM	85.16%
KNN	76.89%

- The accuracy for LR is shown to be 98.675 in Table 1, this is very high but Logistic Regressions are binary in nature and not conducive to more complex problems. Although the results look good, this isn't the most reliable algorithm.
- For SVM, linear kernel is used along with hyperparameter value being C=10000.0, gamma=0.0001, degree=9. Prediction accuracy is 85.16% and scatter plot shows prediction for all the analog drugs. Figure 7 shows the values being concentrated on on the either side of the hyperplane and some values being deviated from the this linear data line. It fails to show the correct categorization of the data values.
  The overall accuracy is not bad but is low enough that there seems that improvements can be made.
- KNN gave an accuracy of 76.89% for the prediction. On visualizing the result Figure 8, the graph seems to be dispersed linearly. This algorithm fails to determine meaningful relationship between the analog drugs. The accuracy is comparatively lower than the previous two algorithms.
- Use of clustering method, hierarchical clustering made improved visualization of the prediction result. Figure 9 shows the data values after clustering. The two clusters seems to be overlapping on one another. Determining the boundary between the the two cluster was difficult this made the analysis task a bit tougher. Hierarchical clustering did not proved to be the best option.
- Figure 10 shows the output of Gaussian Mixture Model. The similar analog drugs are grouped into an elliptical cluster where as linear cluster shows the analogs that are different from the actual inhibitor. This algorithm gives a simulation of stronger cluster between the values.

3 CONCLUSION

After the implementation of the different type of machine learning algorithms, the output in terms of visualization of clustering that unsupervised learning gave a better result than the supervised learning. Usage of the dependent feature of molecular weight gave good prediction for supervised learning overall but showed little clustering compared to unsupervised learning. Logistic regression was exceedingly high at 98.67 but is not the best model to use for complex problems. SVM and KNN had fairly good accuracy but higher accuracy would be desired. Gaussian mixture model which is a type of unsupervised algorithm outperformed all the other machine learning algorithm in terms of grouping similar analogs. This work could be further extended in using other algorithms of supervised and unsupervised learning. Further other data sets featuring analogs of drugs can be taken and see if there is any correlation. And finally, the code of the algorithms tested here can be manipulated further. Testing for different kernels in SMV and using different independent variables for supervised learning.

4 REFERENCE

[1] Manning G, Whyte DB, et al. (2002). "The protein kinase complement of the human genome". *Science*. **298** (5600): 1912–1934.

[2] Carvajal RD, Tse A, Schwartz GK. Aurora kinases: new targets for cancer therapy. Clin Cancer Res. 2006;12(23):6869–6875

[3] Marumoto T, Zhang D, Saya H. Aurora-A – a guardian of poles. Nat Rev Cancer. 2005;5(1):42–50.

[4] Zhou H, Kuang J, Zhong L, et al. Tumour amplified kinase STK15/BTAK induces centrosome amplification, aneuploidy and transformation. Nat Genet. 1998;20(2):189–193.

[5] Nadler Y, Camp RL, Schwartz C, et al. Expression of Aurora A (but not Aurora B) is predictive of survival in breast cancer. Clin Cancer Res. 2008;14(14):4455–4462.

[6] Reiter R, Gais P, Jutting U, et al. Aurora kinase A messenger RNA overexpression is correlated with tumor progression and shortened survival in head and neck squamous cell carcinoma. Clin Cancer Res. 2006;12(17):5136–5141.

[7] Yakushijin Y, Hamada M, Yasukawa M. The expression of the aurora-A gene and its significance with tumorgenesis in non-Hodgkin's lymphoma. Leuk Lymphoma. 2004;45(9):1741–1746.

[8] Gorgun G, Calabrese E, Hideshima T, et al. A novel Aurora-A kinase inhibitor MLN8237 induces cytotoxicity and cell-cycle arrest in multiple myeloma. Blood.

[9] Cervantes A, Elez E, Roda D, et al. Phase I pharmacokinetic/pharmacodynamic study of MLN8237, an investigational, oral, selective aurora a kinase inhibitor, in patients with advanced solid tumors. Clin Cancer Res. 2012;18(17):4764–4774.

[10] Fathi AT, Wander SA, Blonquist TM, et al. Phase I study of the aurora A kinase inhibitor alisertib with induction chemotherapy in patients with acute myeloid leukemia. Haematologica. 2017;102(4):719–727.

[11] Zhang M, Huck J, Hyer M, Ecsedy J, Manfredi M (2009) Effect of aurora a kinase inhibitor MLN8237 combined with rituximab on antitumor activity in preclinical B-cell non-Hodgkin's lymphoma models. J Clin Oncol 27(15_suppl):8553–8553.

[12] Huang, C., Clayton, E. A., Matyunina, L. V., Mcdonald, L. D., Benigno, B. B., Vannberg, F., & Mcdonald, J. F. (2018).

Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Scientific Reports*, *8*(1). doi: 10.1038/s41598-018-34753-5

[13] Hibert, M. & Lopez, P. The world's technological capacity to store, communicate, and compute information. *Science* 332, 60–65 (2011).

[14] Willett, Peter, Barnard, John M. and Downs, Geoffry M. (1998). "Chemical Similarity Searching" (PDF). *Journal of Chemical Information and Computer Sciences*. **38** (6): 983–996.

[15] Willett, Peter, Barnard, John M. and Downs, Geoffry M. (1998). "Chemical Similarity Searching" (PDF). Journal of Chemical Information and Computer Sciences. 38 (6): 983–996. CiteSeerX 10.1.1.453.1788. doi:10.1021/ci9800211.

[16] Application of machine learning techniques to tuberculosis drug resistance analysis.Kouchaki S1, Yang Y1, Walker TM2,3, Sarah Walker A2,3,4, Wilson DJ5, Peto TEA2,3, Crook DW2,3,6; CRyPTIC Consortium, Clifton DA1.

[17] Critical risk-benefit assessment of the novel anti-cancer aurora a kinase inhibitor alisertib (MLN8237): A comprehensive review of the clinical data. Tayyar Y1, Jubair L2, Fallaha S2, McMillan NAJ2.