International Journal of Advance Engineering and Research Development (IJAERD)
Volume 1,Issue 6,June 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

Drug Discovery in Chemoinformatics Using Adaptive Neuro-Fuzzy Inference System

Kirti Kumari Dixit[1], Dr. Hitesh B. Shah[2]
[1]*Information Technology, G.H.Patel College of Engineering, dixit.mahi@gmail.com*
[2]*Electronics & Communication, G.H.Patel College of Engineering, hiteshshah@gcet.ac.in*

Abstract—Drug Discovery is one of the Stages of the Drug Development. It is a major area of application of Chemoinformatics principle. Drug Discovery deals with identifying potential compounds that are drugs from the large chemical space. These Drugs should have relevant features or characteristics to deal with the disease. Drug Discovery is consuming task in the development of Drug. The number of Drugs that are identified today is less as compared to the rate of unknown disease in today's life. Machine Learning Techniques are well known for their supremacy in the area of Classification, Regression, Pattern Recognition, etc problems. In this paper we have tried to enhance the process of Drug Discovery using Machine learning techniques like Neural Network and at the end comparison has been done with a neuro-fuzzy classifier, ANFIS, whose results are remarkably better than the former.

Keywords- chemoinformatics, Drug Discovery, Molecular Descriptor, classification, Artificial Neural Network, Adaptive Neuro Fuzzy Inference System.

I. INTRODUCTION

Chemoinformatics is the term that has been recently coined to represent a discipline that organizes and co-ordinates the application of computers in chemistry. Another well known definition of chemoinformatics is that "It is informatics based approach to solve chemistry problems." Chemoinformatics is composed of two terms: "chemo- means chemistry" and "Informatics- means computers in European countries". Chemoinformatics techniques like Virtual Screening, Virtual Libraries, QSAR/QSPR studies has been applied to drug industry but now it is applied in full range of chemistry. We will focus here on Drug Discovery, one of the emergent applications where principles of chemoinformatics are followed and implemented effectively.

Drug Discovery is the process of identifying innovative leads with potential interaction to specific target [1]. On the other hand Drug Development is the process of making drugs to the market after the series of clinical/non-clinical tests as well as approval by the respective standard authority. This whole process of Drug Discovery & Development takes about 12 to 15 years to complete and costs more than 1 Billion U.S Dollar. This makes the whole process time-consuming and very expensive that cannot be ignored for long time. As the increasing rate of unknown disease can cause adverse effect on humanity, one cannot ignore the above mention problem of Drug Discovery. So, we need some enhance techniques which when applied to the process reduces the time to discover potential drugs. And as the Drug Discovery is crucial problem of Classification, machine learning techniques are well suited for it.

Machine learning techniques are used in the context of chemoinformatics for variety of purpose and functionality. The major functionalities of chemoinformatics where m/c learning techniques are used are chemical structure/property prediction, molecular similarity/diversity analysis, virtual screening, qualitative/quantitative structural/activity/property relationships, ranking chemical structures, representation of chemical compounds/reactions, classification/search/storage methods, high throughput docking, drug discovery, data analysis methods, etc [6].

In this paper, we had applied the two very famous machine learning techniques, Artificial Neural Network and Adaptive Neuro-Fuzzy Inference System in the task of classifying Drug/Non-Drug from

mixed collection of around 5000 instances that comprises Biological compounds, Natural Compounds, chemicals from Drug Bank's Database. The Chemicals are classified on the basis of features mentioned by Lipinski's rule of Drug.

First section of the paper gives the brief introduction about the context. Section two provides the brief overview of Drug Development process, its problems and challenges. Next section gives the stand up for Experimental setup and Simulation results. The last section will end the paper with the fruit of conclusion and references.

II. DRUG DISCOVERY AND DEVELOPMENT STAGES

Drug is nothing but a chemical substance that affects the process of mind and body. Any chemical compound used in the diagnosis, treatment, or prevention of disease or other abnormal condition can be considered as drug. Now, the process of drug discovery aimed at discovering molecules that can be very rapidly developed for effective treatments to meet medical needs. This process generates a large amount of chemical data which is generally referred as information explosion. Due to this it becomes an urgent need to effectively collect, organize, analyze and apply the chemical information in the process of drug discovery and development.

2.1. Traditional Drug Discovery

Traditionally Drug discovery process comprises of following stages-

- *Disease Identification:*

This stage starts with complete understanding of the disease by knowing, how the genes can be altered, how it affects the protein, how this protein will react with each other in living organism, how the affected cells can change the specific tissues and how the disease affects the patient. This stage is also known as pre-discovery phase of the Drug Discovery.

- *Target Identification:*

This stage deals with the identification of protein or gene involved in the disease. That particular protein will be termed as Target. The identified target is separated, crystallized and ligand binding processes are done.

- *Hits Identification:*

This phase deals with the identification of the compounds that likely to bind with the target. It means to find the molecule that can be effective against the target.

- *Pre-Clinical Testing:*

Phase that checks whether the hits can be made into a drug to treat specific disease. Parallel, it is also tested that hits should be not toxic and has minimum side effects.

- *Human Clinical Trial-*

This is the fastest and safest way to find treatments. It acts as the best solution for challenging health disease of human being. Patient with specific disease will be considered for clinical trials. The respective data is collected with respect to time.

- *Approval from the authority and drug in Market-*

In this phase research authorities check the safety and other parameters to approve the drug in the national as well as international market.

It can take up to fifteen years to develop new medicine from the earliest stages of discovery to time it is available in the market for treating patients. The average cost to research and develop each successful drug is estimated to be $800 million to $1 billion. This number includes cost of thousands of failures: For every 5000-10,000 compounds that enter the research and development pipeline, ultimately only one receives approval.

2.2. Modern Drug Discovery and Development

The modern Drug Discovery incorporates two main stages after target identification. These stages are listed below:

- *Lead Identification*

This stage replaces the hit identification stage of traditional process. It includes finding a promising molecule (i.e. lead compound) that could become a drug. In this phase scientists search for the lead compound that can alter the disease course.

- *Lead Optimization*

Lead compounds that survive the initial screening are then "optimized," or altered to make them more effective and safer. By changing the structure of compound, it can gain different properties. Hundreds of different variations or "analogous" of the initial leads are made and tested. The biologist tests the affect of analogous on biological system whereas the chemists take this information to make additional alterations that are then retested by biologists.

Modern Drug Discovery has enhanced the traditional process but it was not as much as effective and efficient as it should be. Still, the Drug Discovery process has to face challenges like determining which compound to be taken from discovery to development process, choice of molecular descriptor (in this case features), classification techniques impose challenges in discriminating between Drug and Non-Drug, chemical database should be updated, and very important complex structures of bio molecules.

Machine Learning techniques can solve the above mentioned problem of chemoinformatics or more specifically Drug Discovery process. Proper feature selection Algorithm like Principal Component Analysis or Genetic Algorithm can be used for the selection of accurate molecular descriptor. Moreover, there are number of ML Techniques are available that can classify the bio molecules into drug/non-drug without imposing any complexities. Here, in next section we are presenting the initial set up for performing our experiment on classifying the chemicals into drugs or non-drugs based on the Lipinski's rule of drug's features. The simulation results are presented after the set up in which comparison has been done between the basic classifier Artificial Neural Network and Neuro-Fuzzy Classifier. The supremacy of neuro-fuzzy classifier over ANN is found to be very interesting and efficient for the Drug Discovery Process.

III. EXPERIMENTAL SET UP AND RESULTS

In order to implement the task of discovering the drug, we need a large set of descriptor and a dataset containing molecules from which a classifier will create two sets of classes, namely Drug and non-Drug.

We use MATLAB R2010b as a simulation tool for classification purpose. The molecular Descriptors are the feature that describes the given molecule based on their topology, chemical composition, behavior of Atoms, etc. The Molecular Descriptor chose by us are taken from the Lipinski's rule for Drugs namely, H-Donor, H-Acceptor, Mg. Vol., LogP. Moreover around more than 5000 instances are collected from DrugBank's Database(25%), PubChem's Database (25%), ChemBL's Database (25%), and Merck's candidates. All the above three mentioned databases are freely publicly accessible.

Moreover, we demonstrated the ANN algorithm for varied number of Samples and the results shows that the Training Performance increases as the no. of samples of Database increases. Results are shown below.

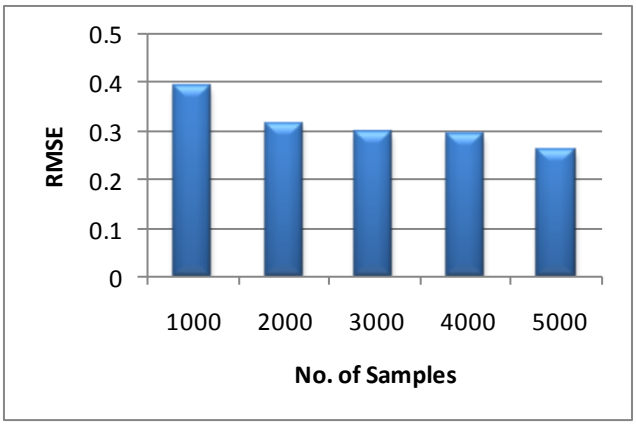

Figure 1: Training Performance w.r.t No. of Samples

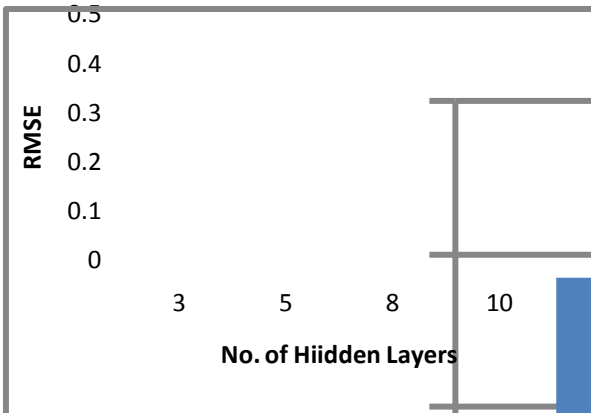

Figure 2: Training Performance w.r.t No. of Hidden Layers

The criterion function is taken as Root Mean Square Error to evaluate the Performance of the Network. Number of Epoch is taken 1000 and number of hidden layers is taken 10 for the above experiment. The next Experiment has been carried out to determine how many hidden layers should be taken in order to have best performance. The Implementation result shows that the Performance of Multilayer Perceptron increases effectively if the number of hidden layer is 10. The figure 2 shows the variations in the Performance of ANN Algorithm with respect to number of hidden layers. The above figure shows that as the number of hidden layer increases the value of Root Mean Square for the Algorithm decreases which results in increasing Training performance. The best Performance of 0.2625 is when the number of hidden layers n Network is 10. So, for carrying out our rest of experiments we took the fixed value of 10 hidden layers in order to get effective results.
The variations of classification Accuracies of training and Testing set by Multilayer perceptron are clearly visualized by the figure 3.

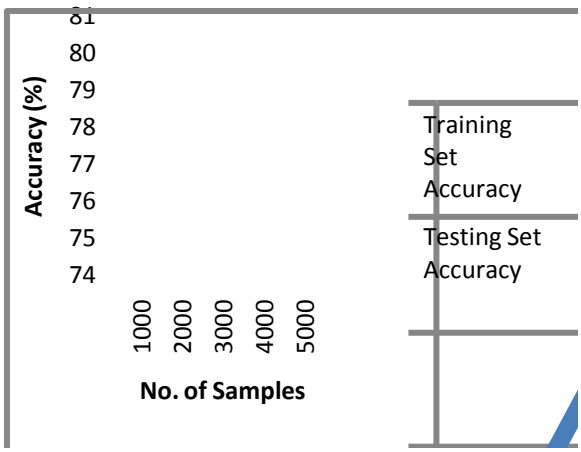

Figure 3: Classification Accuracy for Training and testing set

The Implementation results confirm that classification rate for classifying the chemical dataset into Drug/Non-Drug of Artificial Neural Network is 80% which proves to be much better. But the thing that should be taken into account is that Classification accuracy of testing set is little bit less than that of Training set which can be concern for ANN when there are millions of chemicals to be classified. Moreover, ANN is unable to deal with uncertainty in data which raises a lot of questions on its usability. Therefore, we use Neuro-fuzzy classifier which effectively deals with uncertain data and have the potential advantageous capabilities of neural Network also. The Neuro-Fuzzy Classifier used here is ANFIS which stands for Adaptive Neuro-Fuzzy Inference System. Classification is done using ANFIS and its results are compared with ANN that proves its supremacy in this background.

One thing should be noted that rarely ANFIS is used for Drug Discovery nowadays but if it is implemented as soon as possible in the process, we are effectively benefited from its tremendous performance and accuracy of classification. The results for classification using ANFIS and its comparison with ANN have been shown in Figure 4 and Figure 5 respectively.

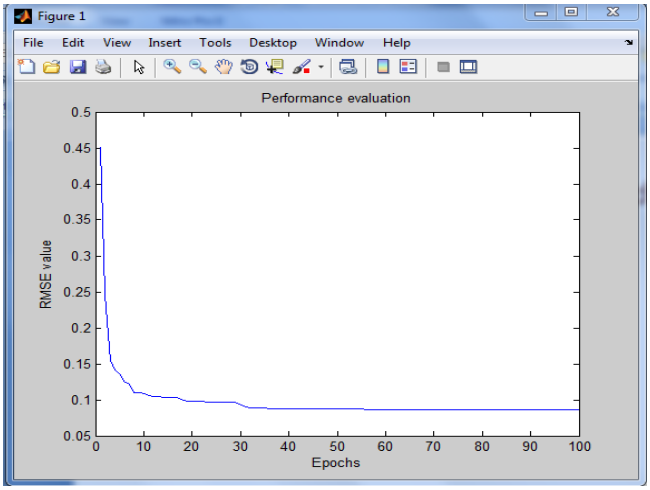

Figure 4: Performance of ANFIS w.r.t Epochs

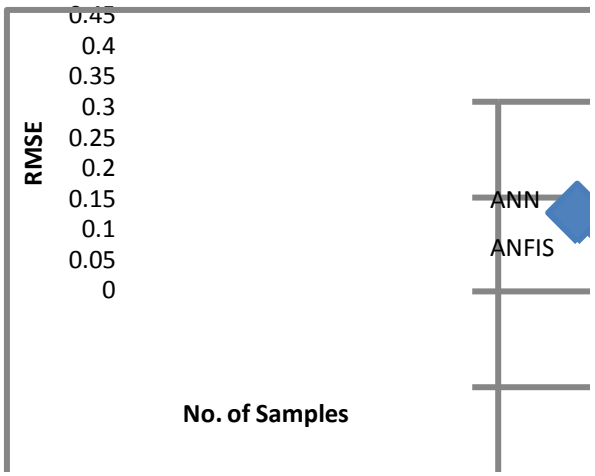

Figure 5: Comparison of Classification Accuracy for ANFIS and ANN

The RMSE of ANFIS and ANN is formulated in tabular form (see Table 1). We can clearly see that MSE of ANFIS for each number of sample is less than that of ANN, thus inferred that better performance of ANFIS than ANN.

Table 1: Comparison of Classification Accuracy for ANFIS and ANN w.r.t number of samples

No. of Samples	ANN	ANFIS
1000	0.391	0.0862425
2000	0.3161	0.090298
3000	0.298	0.0871112
4000	0.288	0.0868264
5000	0.2625	0.0864556

IV. CONCLUSION

The presented ANFIS Model combined the neural network Adaptive capabilities and fuzzy logic Qualitative properties. The classification Accuracy of ANN is 80% and that of ANFIS is 95.94% which is far better. Moreover, the error rate is just 0.08624 in case of ANFIS. Therefore this clearly proves its supremacy in the task.

At last we conclude that ANFIS can be used for the Classification of Chemical Compounds/molecules into Drugs and non- Drugs in order to improve the process of Drug Discovery to a great extent.

REFERENCES

[1]. B.Firdaus Begam, Dr. J.Satheesh Kumar, "A Study on Cheminformatics and its Application on Modern Drug Discovery," *Elsevier*, 2012

[2]. Deepak Bharati, Jagtap RS, Kanase KG, Sonawame SA, Undale VR, and Bhosale AV, "Chemo informatics: Newer Approach for Drug Development," *Asian J. Research Chem*, 2009.

[3]. Evgeny Byvatov, Uli Fechner, Jens Sadowski and Gisbert Schneider, "Comparison of Support Vector Machine and Artificial Neural Network System For Drug/Non-Drug Classification," *J. Chem. Inf. Comput. Sci*, 2003

[4]. V.Arulmozhi, Rajesh Reghunandhan, "Neural Network For Chemoinformatics-A Survey," *ICACCT*, 2013

[5]. Rajesh Reghunandhan, V.Arulmozhi, "Fuzzy Logic for Chemoinformatics-A Review," *journal of Theoretical and applied Information technology*, 2013.

[6]. D. Pugazhenthi, S.P.Rajagopalan, "Machine Learning Technique Approaches In Drug Discovery, Design, Development," *Information Technology Journal* 6(5): 718-724, 2007.

[7]. Vishnu J. Gaikwad, "Application Of Chemoinformatics For Innovative Drug Discovery," *International Journal Of Chemical Sciences And Applications*. Vol. 1, Issue 1; 2010.

[8]. V. Arulmozhi, Reghunandhan Rajesh, "Evolutionary Neural Network For The Classification Of Chemoinformatics Data Sets," *European Journal Of Scientific Research* 2012.

[9]. Axel J. Soto, "On The Use Of Machine Learning Methods For Modern Drug Discovery," *Ai Communications*; 2011.

