Using Supervised Learning and Comparing General and ANTI-HIV Drug Databases Using Chemoinformatics

Taneja Shweta, Raheja Shipra, and Kaur Savneet

Guru Teg Bahadur Institute of Technology
Rajouri Garden, New Delhi- 110064
shweta_taneja08@yahoo.co.in, shipraraheja@gmail.com,
savneetkaur@yahoo.com

Abstract. Earlier People used to discover new drugs either by chance that is serendipity or by screening the natural products. This process was time consuming, costly as well as required a lot of investment in terms of man-hours. The process of discovering a new drug was very complex and had no rational. Prior to Data Mining , researchers were trying computer methods such as potential drug molecules interactions with the targets and that was also time consuming, costly and required high expertise. Data mining is often described as a discipline to find hidden information in a database. It involves different techniques and algorithms to discover useful knowledge lying hidden in the data. Data mining and the term Knowledge Discovery in Databases (KDD) are often used interchangeably. In this paper, we are implementing the classification technique using WEKA tool for the analysis of similarity between GENERAL DRUGS and ANTI-HIV DRUGS.

Keywords: Classification, Chemoinformatics, Data mining, HIV drugs.

1 Introduction

1.1 Motivation

Earlier all over the world, when a new drug was sought pharmacological researchers used to conduct a blind study of tens or hundreds or thousands of chemical compounds, applying them to an assay for a disease. Drug discovery process had no rationale and it tended to be a bit hit and miss.[1].

Then the field of Data Mining emerged. It combines artificial intelligence with innovative knowledge discovery techniques to analyze the results of pharmacological experiments it conducts itself. By relating the chemical structure of different compounds to their pharmacological activity, Now We are able to learn which chemical compounds should be tested next, bringing a degree of predictability to drug screening procedures. This will help scientists and pharmaceutical companies identify more effective compounds to treat different diseases, allowing them to find drug leads in a fraction of the time and at a fraction of the cost of current methods and could minimize the need for random testing of chemical compounds.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 177–183, 2009.

In our work, we are trying to find the general drugs which can be used instead of ANTI-HIV drugs .All available ANTI-HIV drugs in the market have some side effects. The solution of this problem was to either search for a new molecule which could be very expensive or use the already existing drug .The general drugs and their properties are already known and it can be used. Using the WEKA tool, we have implemented the Classification Technique and produced decision Trees[10] of both the databases .There are various other classifiers also available but we selected decision trees for our similarity search as they can be directly converted into the Classification rules. To construct a rule, a path is traced from root to a leaf node.

1.2 ANTI-HIV Drugs

A virus from the Latin, meaning toxin or poison is a sub-microscopic infectious agent that is unable to grow or reproduce outside a host cell. Viruses infect all cellular life. Viruses consist of two or three parts: all viruses have genes made from either DNA or RNA, long molecules that carry genetic information; all have a protein coat that protects these genes; and some have an envelope of fat that surrounds them when they are outside a cell. A retrovirus is a virus with an RNA genome that replicates by using a viral reverse transcriptase enzyme to transcribe its RNA into DNA in the host cell. The DNA is then incorporated into the host's genome by an integrase enzyme. Antiretroviral drugs are medications for the treatment of infection by retroviruses primarily HIV. There are different classes of antiretroviral drugs that act at different stages of the HIV life cycle. Antiretroviral drugs are broadly classified by the phase of the retrovirus life-cycle that the drug inhibits.

1.3 Data Mining

The term ‘data mining’ is the extraction of interesting (non-trivial, implicit, previously Unknown and potentially useful) information or patterns from data in large databases. Its alternative names are: Knowledge discovery (mining) in databases (K.D.D.), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

2 Classification Technique

Data Classification is a supervised learning technique.

Classification is a Two-Step Process:

1. Model construction: It is used for describing a set of predetermined classes
    - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
    - The set of tuples used for model construction: training set

- The model is represented as classification rules, decision trees, or mathematical formula

2. Model usage: It is used for classifying future or unknown objects
   - Estimate accuracy of the model
   - The known label of test sample is compared with the classified result from the model
   - Accuracy rate is the percentage of test set samples that are correctly classified by the model
   - Test set is independent of training set, otherwise over-fitting will occur.

3 Decision Tree

A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. The topmost node in the tree is the root node. Decision trees are used for classification. Given a tuple X,for which the class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple.Decision trees can be easily converted to classification rules. The decision tree classifiers are very popular as it does not require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans .In general decision tree classifiers have good accuracy.

4 About the Tool WEKA

“WEKA” stands for the Waikato Environment for Knowledge Analysis. (Also, the weka pronounced to rhyme with Mecca, is a flightless bird with an inquisitive nature found only on the islands of New Zealand.). Weka is developed at the University of Waikato in New Zealand, the system is written in JAVA an object oriented language that is widely available on all computer platforms, and weka has been tested under various operating systems like Linux, Windows, and Macintosh. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre- and post processing and for evaluating the result of learning schemes on any given dataset. There are several different levels at which Weka can be used. First of all it provides implementations of state-of-the-art learning algorithms that you can apply to your dataset from the command line. It also includes a variety of tools for transforming datasets, like the algorithms for discrimination .We can preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier and its performance—all without writing any program code at all.

5 Dataset Construction

5.1 Drug Bank- A Knowledge Base for Drugs, Drug Actions and Drug Targets

Drug Bank is a richly annotated resource freely available on internet that combines detailed drug data with comprehensive drug target and drug action information. Since its first release in 2006, Drug Bank has been widely used to facilitate in silico drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction and general pharmaceutical education.

We have created two databases with descriptors (Formula weight, Predicted water solubility, Predicted Log P, Experimental Log P, and Predicted Log S) as follows-

- A section of 350 drugs has been made from DRUG BANK
- Then a database of 25 ANTI HIV drugs is made.

A sample of records containing 10 instances is as shown in following tables-Table 5.1 and Table 5.2.

Table 5.1. Database of General drugs

Name	Type	Formula Weight	Predicted Water Solubility mg/mL	P LogP	E LogP	P LogS
Astrozole	Approved	293.3663	6.61E-02	2.32	2.4	-3.65
Acamprosate	Approved	181.21	1.88E+01	-1.78	-1.1	-0.98
Acetazolamide	Approved	222.245	2.79E+00	-0.39	?	-1.9
Adenosine	Approved	267.2413	1.40E+01	-1.2	-1.6	-1.28
Alendronate	Approved	249.096	1.69E+01	-1.34	-4.3	-1.17

Table 5.2. Database of Anti HIV Drugs

Name	Type	Formula Weight	Predicted Water Solubility mg/mL	P LogP	E LogP	P LogS
Abacavir	Approved	286.3323	1.21E+00	0.61	1.1	-2.37
Amprenavir	Approved	505.627	4.91E-02	1.85	?	-4.01
Atazanavir	Approved	704.8555	3.27E-03	4.08	4.5	-5.33
Cidofovir	Approved	279.187	1.15E+01	-2.11	-3.9	-1.38
Darunavir	Approved	547.664	6.68E-02	1.89	1.8	-3.91

6 Experimental Results

There are various tools which are used for Data Mining .Because the WEKA tool is a freely available tool, that's why we implemented classification technique with WEKA tool. WEKA supports input files in 2 formats- ARFF format and CSV format. The figures 6.1 and 6.2 given below show the decision trees of GENERAL drugs and ANTI-HIV drugs.

6.1 Decision Trees

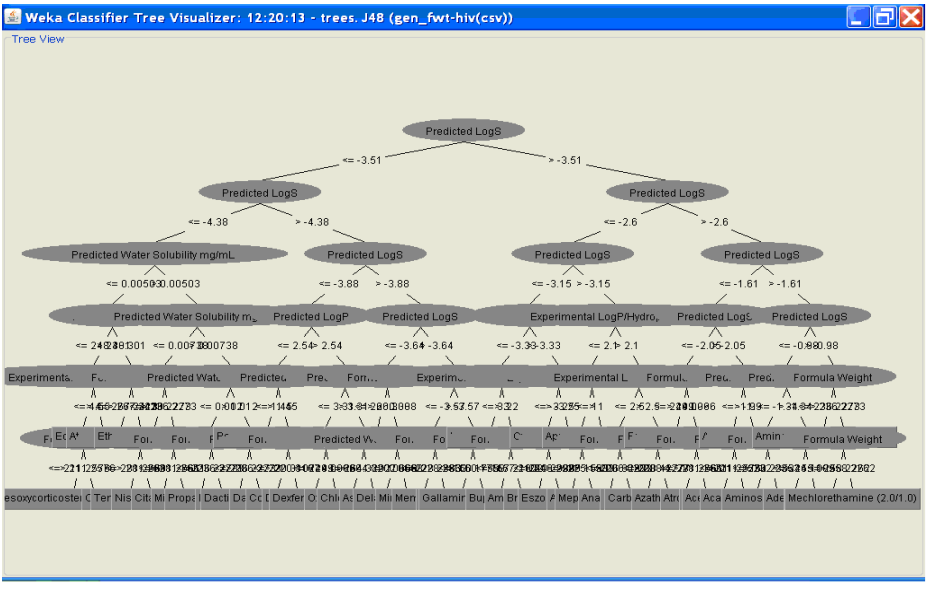

Fig. 6.1. Decision tree of general drugs

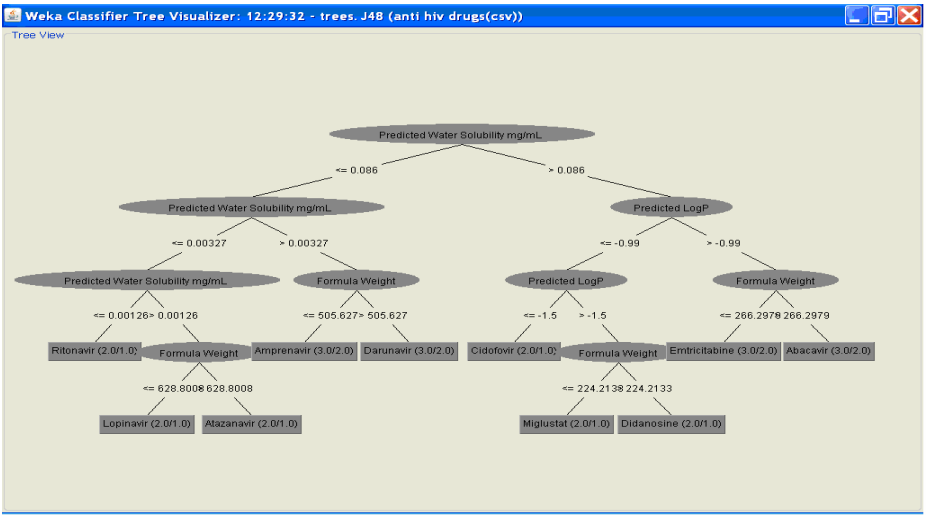

Fig. 6.2. Decision tree of Anti HIV drugs

6.2 Classification Rules

GENERAL DRUGS

1. if P.W.S<=0.00503 and F.W.T<=248.301 and F.W.T<=211.2576
   =>Drug=**Celecoxib**
2. if PWS<=0.00503 and FWT<=248.301 and FWT>211.2576
   =>Drug=**Desoxycorticortesene Pivalate**
3. if PWS<=0.00503 and FWT>.248.301 and FWT<267.2413
   =>Drug=**Atonoxetine**
4. if PWS>0.00503 and PWS>0.00738 and PWS<=0.012 and
   FWT>236.2273=>Drug=**Propafenone**
5. if PWS>0.00503 and PWS>0.00738 and PWS<=0.012 and FWT>236.227
   =>Drug=**Dactinomycin**

ANTI HIV DRUGS

1. if PWS<=0.086 and PWS<=0.00327 and PWS<=0.00126=>Drug=**Ritonavir**
2. if PWS<=0.086 and PWS<=0.00327 and PWS>=0.00126 and
   FWT<=628.800=>Drug=**Lopinavir**
3. if PWS<=0.086 and PWS<=0.00327 and PWS>=0.00126 and
   FWT>628.800=>Drug=**Atazanavir**
4. if PWS<=0.086 and PWS>0.00327 and
   FWT<505.627=>Drug=**Amprenavir**
5. if PWS<=0.086 and PWS>0.00327 and
   FWT>=505.627=>Drug=**Darunavir**

7 Conclusion

The table 7.1 given below shows the predicted results of similarity between GENERAL DRUGS and ANTI-HIV DRUGS on the basis of range of descriptors**.**

Table 7.1. Predicted results of similarity between GENERAL DRUGS and ANTI-HIV DRUGS on the basis of range of descriptors

Name	Type	Formula Weight	Predicted Water Solubility mg/mL	P LogP	E LogP	P LogS
Abacavir	Approved	286.3323	1.21E+00	0.61	1.1	-2.37
Amprenavir	Approved	505.627	4.91E-02	1.85	?	-4.01
Atazanavir	Approved	704.8555	3.27E-03	4.08	4.5	-5.33
Cidofovir	Approved	279.187	1.15E+01	-2.11	-3.9	-1.38
Darunavir	Approved	547.664	6.68E-02	1.89	1.8	-3.91

FWT-Formula Weight.
PWS-PredictedWater Solubility.

As shown in the above table, some ANTI-HIV drugs are similar to GENERAL drugs on the basis of range of descriptors specified. These results are only forecasted results and may be clinically tested.

References

1. Data Mining Promises To Dig Up New Drugs, Science Daily (February 3, 2009)
2. Zimmermann, A., Bringmann, B.: CTC - correlating tree patterns for classification. In: Fifth IEEE International Conference on Data Mining, Novemmber 27-30, p. 4 (2005), doi:10.1109/ICDM.2005.49
3. Bjorn, B., Albrecht, Z.: Tree Decision Trees for Tree Structured Data. Institute of Computer Science, Machine Learning Lab, Albert-Ludwig-University Freiburg, Georges-Kohler-Allee 79, 79110 Freiburg, Germany
4. Ósk, J.S., Steen, J.F., Brunak, S.: Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates 21(10), 2145–2160 (2005), doi:10.1093/bioinformatics/bti314
5. Lumini, A., Nanni, L.: Machine Learning for HIV-1 Protease Cleavage Site Prediction. Elsevier Science Inc., New York (2005)
6. Niko, B., et al.: Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype. IEEE Intelligent Systems 16(6), 35–41 (2001)
7. Hiroto, S.: Mining complex genotypic features for predicting HIV-1 drug resistance. National Institute of Informatics (2007)
8. Lin Ray, S.: A combined data mining approach for infrequent events: analyzing HIV mutation changes based on treatment history. Stanford University Stanford, CA 94305, United States
9. Wasim, A.M., Sidhu, M.: Chemoinformatics: Principles and Applications. Pesticide Residue Laboratory, Department of Agricultural Chemicals, Department of Agricultural Chemistry and Soil Science, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur-741252, Nadia, West Bengal, India
10. Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., Selbig, J.: Identifying drug resistance-associated patterns in HIV genotypes