International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

Exploring Mutational Pathways of HIV Using Genetic Algorithm

K. M. Monica[1]

Research Scholar, Department of CSE,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

Abstract: ***The Human Immunodeficiency virus (HIV) is caused due to the failure of the immune system that leads to life threatening infections and cancer. Even though most drug resistance mutations have been identified throughout, the dynamics and temporal patterns of these mutations can still be explored. In order to explore the drug resistance mutation, Temporal Bayesian Networks (TNBN) algorithm is used where data is extracted from Stanford HIV drug resistance database. TNBN work needs more iterative process and training data for accurate information and it has failed to compare two different mixture treatments along with the temporal occurrence of drug resistant mutations, in order to predict the most effective treatment. The proposed work performs P_GA (Prediction Genetic Algorithm) which is a prediction scheme with the slotted training dataset changing values. The study shows the proposed techniques provides better results than existing temporal node Bayesian network scheme in terms of accuracy.***

Keywords: Drug resistance, Dynamic Bayesian, Mutation, P-GA, TNBN

1. Introduction

In the world population, it is estimated that 35.3 million of people lives with HIV. The Antiretroviral therapy is highly active and it has led to a significant reduction in AIDS-related morbidity and mortality. The treatments used for HIV helped many people lead longer, healthy lives. However, the side effects caused by the drugs of AIDS and HIV are also common. Although most drug resistance mutations have been well identified, the appearance of temporal patterns of these mutations can still be further explored. And the evaluation of these treatment suggestion schemes should be well defined. The use of models to predict best treatment and pathways as well as enhanced temporal patterns of appearance of adaptive mutations could greatly benefit clinical management of individuals under antiretroviral therapy. The goal of the work is to test whether the algorithm could predict the mutational pathways for specific drugs. In the present study the problem of finding mutation-mutation and drug-mutation associations in treatment using the Temporal Nodes Bayesian Networks is addressed.

The enhanced work plans are to compare the different treatments along with the temporal changes of drug resistant mutations, to predict the most effective treatment. The rest of the report is as follows. In Section 2, literature has been reviewed. Section 3 gives the problem formulation. Section 4 describes the Temporal Bayesian networks algorithm. Section 5 describes the proposed work using Genetic algorithm and association rules. Section 6 gives the results and analyses of the algorithm Section 7 concludes the work.

2. Literature Review

2.1 Classifying HIV Status using Radial Basis Function

Neural Networks are used as pattern recognition tools in data mining to classify HIV status of individuals based on demographic and socio-economic characteristics. The radial basis function (RBF) neural network architecture was used for this study since as preliminary design showed this architecture to be the most optimal. The Bayesian method used was approximated with the evidence framework. The design of classifiers involves the assessment of classification performance, and this is based on the accuracy of the prediction using the confusion matrix [1].

2.2 Using Logistic regression and Bayesian networks to predict the stroke of HIV

Predictive models of stroke in these populations can helps planning targeted strategies to reduced morbidity and mortality in HIV/AIDS populations. Three different statistical models were used: multivariate logistic regression (LR), Bayesian networks (BN) and a combination of both. Goodness of fit was evaluated with the area under the curve and reliability was tested with three-cross validation. Using BN analysis, 75% of stroke cases were accurately predicted. Using a combination of LR and BN, only 72% of the cases were correctly classified [3].

2.3 Genotypic estimation of transmitted HIV using calibrated population resistance tool

The calibrated population resistance (CPR) tool is a web-accessible program for performing standardized genotypic estimation of transmitted HIV-1 drug resistance. Although designed specifically for surveillance of HIVDR, it has proposed that the framework implemented in the CPR program represents a prototype for other areas of molecular epidemiology in particular studies of microbial drug resistance and hence the estimation is not successful [5].

2.4 Temporal Bayesian Network Models for Analyzing HIV Mutations

Evolution is an important aspect of viral diseases such as influenza, hepatitis and the human immunodeficiency virus (HIV). In this paper they use two approaches for modeling the relationships between antiretroviral drugs and HIV mutations, in order to analyze temporal occurrence of specific mutations in HIV that may lead to drug resistance.

They compare the strengths and limitations of each of these two temporal approaches for this particular problem and show that the obtained models were able to capture some mutational pathways [9].

2.5 Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries:

Genetic research is increasingly turning to studies of sequence variation in genes encoding proteins of known structure and function. The proliferation of published sequence data and the growth in the number of publications is a boon to this research, but also makes it difficult to keep track of what is known about a gene. The development of public databases and their associated applications will encourage researchers to adopt standard approaches for reporting biological data and, in doing so, new findings will be more effectively subjected to peer review and to subsequent analyses and use by others.

3. Problem Formulation

There are some clear limitations to the existing Bayesian Node network models. While Bayesian models are a useful way to model expert knowledge, it may be difficult to get experts to agree on the structure of the model and the nodes that are important to be included.

3.1 Bayesian networks

These are probabilistic graphical models particularly well suited to deal with uncertainty that represents a set of random variables. It is a visual representation, a graph consisting of nodes and edges facilitating their analysis and interpretation.[1]. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptom. Given symptoms, the algorithm can be used to compute the probabilities of the presence of various diseases. This type of network is common when only the links between the consecutive stages are allowed. If temporal changes occur infrequently the representation of BN becomes unnecessary. Therefore the alternative technique applied is TNBN (Temporal Nodes Bayesian Networks).

3.2. TNBN Algorithm

A TNBN is a type of PGM (Probabilistic graphical models) in which each node represents an event, and each edge in the graphical structure represents a temporal probabilistic relation. TNBNs are composed by two types of nodes**: instantaneous and temporal** Instantaneous nodes model events in which no time delay is seen before their occurrence, that is, once a parent event takes place the manifestation of the corresponding child event is immediate. Unlike an instantaneous node, a temporal node models the possible time delays between the occurrence of the cause and the observing of the effect. Each temporal node consists of a set intervals in which an event may happen. A "Default" value indicates that the event does not occur. It is important to point out that all root nodes of a TNBN must be instantaneous.

Here is an example of a TNBN that models the event of a collision is provided in Figure. [3], [1]

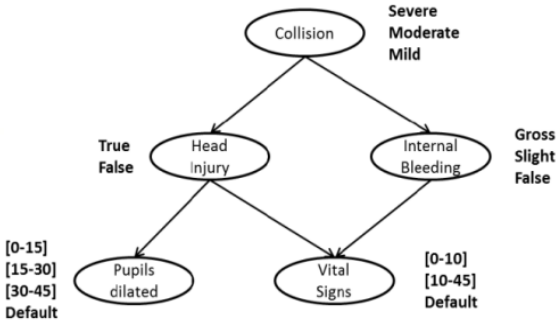

Figure 1: TNBN model.

Figure 1: A Temporal Nodes Bayesian Network modeling the event of a collision. Here, "Collision", "Head injury" and "Internal bleeding" are instantaneous nodes, while "Pupils dilated" and "Vital signs" are temporal nodes each with a set of associated intervals and a default value.

TNBN work needs more iterative process and training data for accurate information. The existing work suffers from the followings disadvantages. Despite the compared models obtained important information they were not capable of providing a global and complete model showing the temporal process of the problem.

Drawbacks:

- Failed to compare two different mixture treatments along with the temporal occurrence of drug resistant mutations, in order to predict the most effective treatment.
- Verification of accuracy against the TNBN system.
- Only provides approximate solution for treatment selection.

4. Proposed Work

A. Implementation of Genetic Algorithm

Genetic algorithms employ metaphor from biology and genetics to iteratively evolve a population of initial individuals to a population of high quality individuals, where each individual represents a solution of the problem to be solved and is compose of a fixed number of genes. In every generation, three basic operators of genetic algorithm i.e. selection, crossover and mutation are applied to each individual. The proposed system implements genetic approach and a P-GP algorithm and extends the perspective of that approach in order to be able to deal with groups, or sub category, of symptoms based on the individuals.

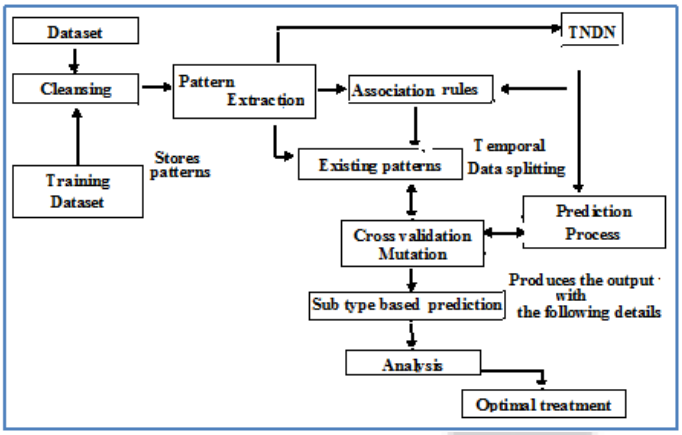

Figure 2: System Architecture

P-GP Algorithm:

Step 1: Generate base conditions:
a) Get patient and treatment data's and associations from the PET model
b) Threshold(maximum value)
c) If the data is categorical- add to the label
d) Else if attribute perform the following
   a. Based on the value it captures least possible symptom from the data
   b. Repeat until checking complete all objects

Step 2: Combine Conditions
Step 3: Finally P-GP algorithm generates best treatment of data using genetic approaches by performing additional crossover and mutation process
a. Read set of patient data attributes
b. Declare a variable to store the output
c. Set the condition (patient dataset, support, confidence, crossover, threshold, mutation value)
d. For each attribute set condition combination
e. Combine conditions (patient dataset, support, confidence, crossover, threshold,
f. Proceed results

B. Implementation of Association rule mining

The next process of the proposed implementation is to mine association rules from the web usage log file. After successful preprocessing the data will be allocated to the next phase, which is known as association rule finding. Two rules are applied for implementing. The first number is called the **Support** for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. The other number is known as the confidence of the rule. **Confidence** is the ratio of the number of transactions that include all items in the consecutive phase.

Apriori Algorithm (1)
Apriori algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules.

Apriori Algorithm (2)

Uses a **Level-wise search**, where **k-itemsets** (An itemset that contains k items is a **k-itemset)** are used to explore (k+1)-itemsets, to mine frequent itemsets from transactional database for Boolean association rules.

First, the set of frequent 1-itemsets is found. This set is denoted L1. L1 is used to find L2, the set of frequent 2-itemsets, which is used to fine L3, and so on, until no more frequent k-itemsets can be found.

Association rule mining process
- Find all **frequent itemsets**:
- Each support **S** of these frequent itemsets will at least equal to a pre-determined min_sup (An itemset is a subset of items in I, like A)

Generate **strong association rules** from the frequent itemsets:These rules must be the frequent itemsets and must satisfy min_sup and min_conf.

The proposed system performs the following steps.
- Predicting sub types.
- Calculates support and confidence of symptoms, treatments and
- Combinational treatment finding.

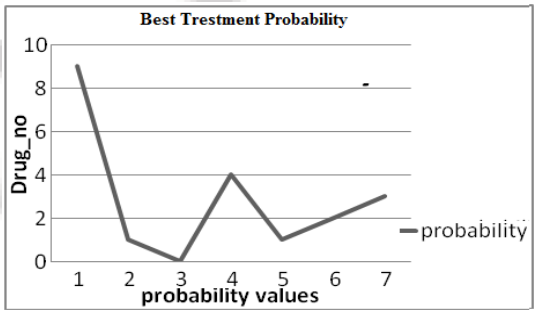

Figure 3: Best Treatment Probability

Advantages:
- Performs the cross validation operation with the mixture of treatments.
- Finds best treatment.
- No need of more training data.
- Performs from the existing report.
- Deals with the several attributes.

The efficiency can be improved and the best treatment can be evaluated by using the Genetic algorithm where the association rules are applied for finding the support and the confidence of the patient treatment. By calling the function of Genetic Algorithm various treatments can be compared by using P-GA (Prediction Genetic Algorithm), the efficient treatment can be provided in terms of accuracy.

5. Simulation Environment

5.1 Dataset description

The datasets have been extracted from HIV drug resistance database. In this project, from the database 15 drugs have been taken where 2 types of drugs are taken which is further divided as 15 drugs. According to the symptom of each patient the 15 types of drugs were given. The symptoms of the patients are divided into 16 attributes and the patients details are classified into 4 attributes along with the 16

symptoms. The graph is evaluated by using the temporal changes that are taking place within a week or a month. Hence various are algorithm used for comparing the treatments that are applied to the patients and hence the efficient treatment can be found out.

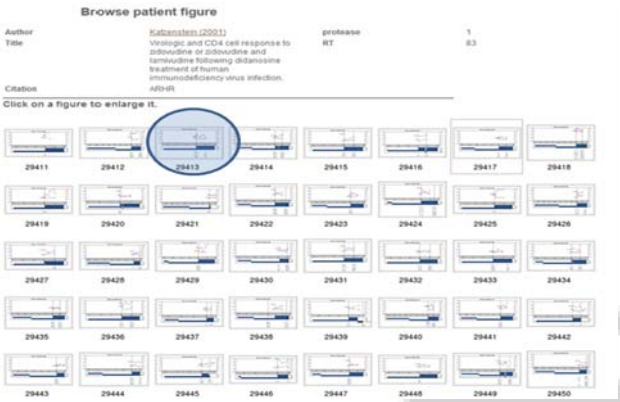

Figure 4: Data set Collection

6. Results and Analysis

In order to evaluate the models and to measure the efficiency in terms of accuracy both the algorithms are compared and partial result has been obtained to give an efficient treatment. The support and confidence are calculated using the association rules where the frequently used drugs can be found out in order to obtain the result. The proposed work they can explore the solution space in multiple directions at once. The algorithm is applied in such a way it can analyze the patient history for further treatment. T he obtaining result can be able to present the model as an interesting tool to explain how mutations interact with each other, providing information not only in the association patterns, but also in the temporal order of appearance of mutations selected by antiretroviral drugs.

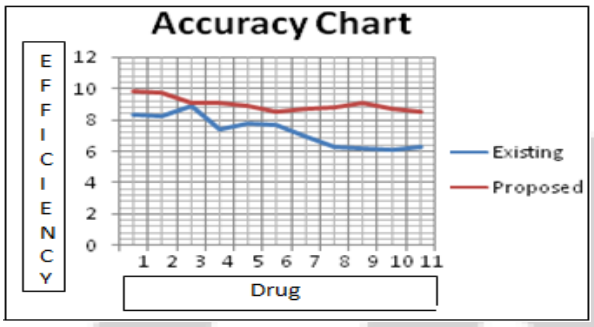

Figure 5: Accuracy Chart

7. Conclusion

The main contribution of this paper is to use a temporal probabilistic approach to understand HIV mutations. The models were urbanized using data and some important known correlated mutations were discovered, as well as other temporal relations. The TNBN approach is compared with other models such as Genetic algorithm and association rules. Though association rules could obtain results similar to the TNBN model, they have several drawbacks. Hence Genetic Algorithm is applied along with this rule to predict the efficient treatment from analyzing the patient history. Only a partial output is obtained and hence the study is still in process to obtain the result in an efficient way. If sufficient data becomes available, social factors including adherence to treatment and drug availability, as well as genetic factors modulating the dynamics could be incorporated to the world. In contrast, the compared model presents a global and easy to understand model that can provide useful temporal information.

References

[1] Prediction of HIV Status from Demographic Data Using Neural Networks
[2] Brain Leke-Betechuoh, Member, IEEE, Tshilidzi Marwala, Member, IEEE, Taryn Tim, Lagazio.
[3] Multilayer Perceptron Network in HIV/AIDS Application *Manaswini Pradhan, Ranjit Kumar Sahu.*
[4] Stroke prediction in a sample of HIV/AIDS patients: Logistic regression, Bayesian networks or a combination of both?
[5] J. Gutierrez a, and C. Yoo b
[6] Bayesian network analysis of resistance pathways against HIV-1 protease inhibitors
[7] K. Deforche a,*, R. Camacho b, Z. Grossman c, T. Silander d, M.A. Soares
[8] Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries
[9] ROBERT W. SHAFER1, DUANE R. JUNG1 & BRADLEY J. BETTS2.
[10] Improving clinical record visualization recommendations with Bayesian stream learning.
[11] Pedro Pereira Rodrigues1;2, Cl_audia Dias2;3, and Ricardo Cruz-Correia2;3
[12] Mo H, King M, King K, Molla A, Brun S, Kempf D. Selection of resistance in protease inhibitor experienced, human immunodeficiency virus type 1 infected subjects failing Lopinavir and Ritonavir based therapy: mutation.
[13] Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, et al. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proceedings of the
[14] National Academy of Sciences of the United States of America 2002;99(12):8271–6.
[15] The calibrated population resistance tool: standardized genotypic estimation of transmitted HIV-1 drug resistance
[16] Robert J. Gifford1,*, Tommy F. Liu1, Soo-Yon Rhee1, Mark Kiuchi1, Stephane Hue2,Deenan Pillay2,3 and Robert W. Shafer1
[17] Galán S, Arroyo-Figueroa G, Dí ez F, Sucar L. Comparison of two types of event Bayesian networks: a case study. Applied Artificial Intelligence 2007;21(3):185.
[18] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. Journal of Computational Biology 2000;7(3-4):601–20.
[19] Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, et al. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype.

Proceedings of the National Academy of Sciences of the United States of America 2002;99(12):8271–6.

Author Profile

K. M. Monica has completed her B.E Computer Science and Engineering from C.I.E.T College of Engineering and Technology, Coimbatore and is pursuing M.E Computer Science and Engineering from Avinashilingam Institute for Home Science and Higher Education for Women University.