A COMBINED DATA MINING APPROACH FOR INFREQUENT EVENTS: ANALYZING HIV MUTATION CHANGES BASED ON TREATMENT HISTORY

Ray S. Lin[1*], Soo-Yon Rhee[2], Robert W. Shafer[2], and Amar K. Das[1]

[1]Stanford Medical Informatics and [2]Division of Infectious Diseases
Department of Medicine, Stanford University
Stanford, CA 94305, United States
[*]Email:raylin@stanford.edu

Many biological databases contain a large number of variables, among which events of interest may be very infrequent. Using a single data mining method to analyze such databases may not find adequate predictors. The HIV Drug Resistance Database at Stanford University stores sequential HIV-1 genotype-test results on patients taking antiretroviral drugs. We have analyzed the infrequent event of gene mutation changes by combining three data mining methods. We first use association rule analysis to scan through the database and identify potentially interesting mutation patterns with relatively high frequency. Next, we use logistic regression and classification trees to further investigate these patterns by analyzing the relationship between treatment history and mutation changes. Although the AUC measures of the overall prediction is not very high, our approach can effectively identify strong predictors of mutation change and thus focus the analytic efforts of researchers in verifying these results.

1. INTRODUCTION

Many databases contain a large number of variables, among which events of biological relevance could be rare and difficult to predict accurately. Combining different data mining methods may be effective in discovering the relationship between rare events (e.g., mutation changes in genotype-test results) and other measured factors (e.g., phenotypic or environmental variables). In our work, we have investigated the relationship between mutation changes in the HIV protease gene and information on patients' recent antiretroviral treatment, using data from the HIV Drug Resistance Database (HIVDB) at Stanford University (hivdb.stanford.edu)[1]. In this paper, we present an evaluation of an approach combining association rule analysis, logistic regression, and classification trees to predict the occurrence of mutations in the HIV protease gene based on treatment history.

Association rule analysis is a popular method for mining commercial transaction databases[2] and has been explored in finding patterns in biomedical data, such as gene regulatory elements in microarray data[3], gene expression and co-regulated clusters[4], and protein-protein interactions[5]. Association rule analysis is an efficient method to discover relations hidden in sparse large database[6]. However, when the events of primary interests are rare in the database, the rules being mined may have very low support (<5%) and cannot reach confident conclusion about the associations. Therefore, other analytical methods are necessary in order to conduct more elaborate investigation about the associations being discovered.

Logistic regression is a classic method for predicting binary outcomes and identifying risk factors in clinical research[7]. Classification trees are powerful analytical tools that produce interpretable results (tree diagrams) and thus have been widely used in predicting various clinical and biomedical events. In HIV research, it has been used to analyze the association of antiretroviral resistance mutations with response to therapy[8] and to predict drug resistance based on HIV mutations[9]. While logistic regression assumes the linearity and independence of the predictors, classification trees can be used to discover the interactions among predictors that do not exhibit strong marginal effects[10]. The classification trees explore high-order effects by the nature of recursive partitioning[2]. However, both of these two methods require pre-specified outcome variables and

predictors and are not effective in scanning through a sparse database with a large number of variables.

Combing these three methods may potentially mitigate each of their limitations. Studies have shown the effectiveness of using different methods in classification[11,12] and the combination of association rule analysis with classification methods[13,14]. However, only few studies explored the combinatorial usage in biomedical domain[15].

In this study, we have undertaken the following approach: We first use association rule analysis to scan through the occurrences of HIV protease gene mutations and identify potentially interesting patterns with relatively high frequency. Then logistic regression and classification trees are used to further investigate these patterns by analyzing the relationship between treatment history and mutation changes. We have found that, whereas association rule analysis can effectively focus interesting patterns for further investigation, logistic regression and classification trees can potentially identify both the linear and high-order relationships in the database.

2. BACKGROUND

Significant research efforts have been undertaken in investigating the association between HIV gene mutations and antiretroviral therapy. Mutations on the protease and reverse transcriptase genes—targets of antiretroviral drugs—have been shown associated with drug resistance[16]. On the one hand, these mutations can cause treatment failure; thus, it is important to be able to predict drug resistance based on the specific HIV mutations and identify the best treatment for the patients [8]. On the other hand, the mutation change may also be the result of certain antiretroviral treatments; therefore, it is also crucial to predict the mutation change based on patients' treatment history.

There have been a number of studies examining the occurrence of HIV gene mutations after initial exposure to specific antiretroviral treatments[16]. Few have investigated sequential mutation changes in patients changing antiretroviral drugs in the context of comprehensive drug treatment histories, such as those available in HIVDB. With data from over 2000 subjects in HIVDB, we undertook a combined data-mining approach to predict mutation changes in the HIV protease gene (P1 to P99) based on antiretroviral drug history.

3. METHOD

We derived a dataset from HIVDB that included 2,681 unique patients who had more than one HIV protease genotype-test result and who had treatment history recorded during "time windows," starting with one genotype test result and ending with another measure. For each of the 99 coding positions in the HIV protease gene, the occurrence of a mutation change is identified as a difference between the test result at the beginning of the time window and that at the end. Antiretroviral drugs administered within the time window are considered to be the predictors of mutation change. Each of the 7 protease inhibitors (PIs)—abbreviated APV, IDV, NFV, RTV, SQV, LPV, and ATV—is represented as an individual predictor. Nucleoside reverse transcriptase inhibitors (NRTI) and non nucleoside reverse transcriptase inhibitors (nNRTI) are treated as two aggregated predictors.

We first utilized association rule analysis to scan through the database for identifying patterns (i.e., pairs of predictors and mutation changes) of relatively high frequency. In order to achieve high sensitivity in identifying potentially interesting patterns, rules are mined by the *a priori* algorithm using minimum support 0.002 and minimum confidence 0.1. Each of the coding position identified in the association rules was further analyzed by a logistic regression model and a classification tree. For each model, the dependent variable is a binary variable indicating whether there is a mutation change at that coding position; the independent variables are the 9 treatment predictors. In classification trees, Gini index of diversity is used as the splitting criterion. The trees are pruned back by 1-SE rule based on 10-fold cross validation error rates[2]. We evaluated the performance by area under ROC curve (AUC) analysis, and assessed the sensitivity and specificity at the optimal point.

4. PRELIMINARY RESULTS

4.1. Descriptive Statistics

In average, the length of the time window is 391 days, the number of mutation changes per time window is 3.01, and the percentage of mutation changes per coding position is 3%. The occurrence of each PI drug in a single time window ranged on average from 1% (ATV) to 20% (IDV); the occurrence of NRTI and nNRTI were 63% and 33% respectively.

4.2. Association Rules Mining

In total, 449,077 rules are mined in association rule analysis. Among them, we examine further 1,406 rules with treatment on the left hand side and mutation change on the right hand side. The highest support is

0.04 and highest confidence is 0.50. Table 1 shows example mined association rules that have relatively high confidence and support.

4.3. Genotype Change Prediction

We identified nine unique coding positions from the 50 rules with highest confidence and the 50 rules with highest support. Each of these nine positions was analyzed by a logistic regression model and a classification tree. The models predict whether there is a mutation change at a particular position based on the treatment. The AUC measurement and sensitivity/ specificity at the optimal cut point are summarized in Table 2. The AUC measurements show that the overall prediction performance is not high (from 0.56 to 0.70). However, both methods identify strong predictors of mutation change in each of the nine coding positions. In average, logistic regression identifies four strong predictors (with $p < 0.001$), and classification trees identify drug combinations (consisting of two to seven drugs) as predictors of the mutation change at each position.

Table 3 shows the effects of treatment on one particular mutation change (P54) estimated by logistic regression. In this analysis, APV, IDV, RTV, SQV, and LPV are strong risk factors for developing this mutation change. NRTI and nNRTI show strong protective effect while other drugs are not associated with the mutation change.

Table 1. Association rules with relatively high confidence (Conf.) and support (Supp.)

Rule	Supp.	Conf.
{APV, SQV, LPV}⇒ {P10}	0.002	0.50
{APV, RTV, LPV, nNRTI}⇒ {P10}	0.003	0.47
{APV, SQV, LPV}⇒ {P54}	0.002	0.45
{IDV} ⇒ {P46}	0.039	0.20
{RTV}⇒ {P71}	0.038	0.22

Table 2. The prediction performance of classification trees and logistic regression indicated by AUC and by sensitivity and specificity, in parentheses

Coding Position	Classification Trees	Logistic Regression
P10	0.64 (0.51, 0.69)	0.61 (0.60, 0.55)
P13	0.62 (0.67, 0.53)	0.60 (0.68, 0.50)
P20	0.65 (0.62, 0.64)	0.65 (0.66, 0.57)
P36	0.63 (0.66, 0.55)	0.63 (0.66, 0.55)
P46	0.67 (0.63, 0.62)	0.64 (0.57, 0.64)
P54	0.62 (0.54, 0.76)	0.70 (0.64, 0.64)
P71	0.62 (0.60, 0.58)	0.60 (0.51, 0.65)
P82	0.63 (0.58, 0.70)	0.67 (0.57, 0.68)
P90	0.56 (0.46, 0.74)	0.64 (0.56, 0.65)

Table 3. The treatment effects on the mutation change of P54 estimated by logistic regression

Treatment	Odds Ratio (95% CI)	p value
APV	4.72 (3.55, 6.28)	<0.001
IDV	1.56 (1.23, 1.97)	0.003
NFV	0.96 (0.73, 1.25)	0.75
RTV	1.77 (1.36, 2.30)	<0.001
SQV	1.68 (1.27, 2.21)	<0.001
LPV	2.45 (1.81, 3.31)	<0.001
ATV	0.95 (0.36, 2.47)	0.91
NRTI	0.65 (0.51, 0.83)	<0.001
nNRTI	0.66 (0.52, 0.82)	<0.001

5. DISCUSSION

In this study, we analyze data on mutation changes stored in the HIVDB and investigate their association with patients' history of antiretroviral treatment. In the HIVDB, the occurrence of mutation changes in the HIV protease gene is less than 15% and the frequency of the combination of these mutation changes and contextual antiretroviral history is less than 5%. No single data-mining method may adequately find predictors of mutation change. Therefore, we use a novel combined approach to analyzing the database. Association rule analysis is first applied to the database for identifying patterns with relatively high support and confidence. Logistic regression and classification trees are then used to predict the mutation change in specific coding positions.

Although the AUC measurements resulting from our approach do not show high overall prediction performance, we can effectively identify strong predictors of mutation change. A preliminary analysis of these results validates their relevance based on prior studies. The rules listed in Table 1 are consistent with previous research in the association between treatments and HIV mutations[16]. P10 is a well-known polymorphism position. It is not surprising that the mutation change in P10 is associated with different treatments. APV, SQV, and LPV are all shown associated with P54 (specifically I54V) mutation in previous research. The association rule shows a similar association between these three drugs and mutation change in P54. Similarly, IDV has been shown associated with P46 (M46I and M46L) mutation and P71 has been shown associated with P71 (A71V) before, and the rules also show these associations in the mutation change.

With the P54 mutation changes, past studies have shown associations between RTV and I54V/I54L; IDV and I54V; NFV and I54V/ I54L; ATV and I54L. The results of logistic regression find similar

associations in RTV and IDV but not in NFV and ATV (Table 3). One limitation of our study is that we did not distinguish the specific genotype mutation in each change, since these events are too rare to be analyzed with such differentiation. As a result, we were not able to compare our data mining results with existing literature on genotype-specific mutation changes. In addition, this study did not distinguish the treatment-naïve patients from the patients changing different drugs during the treatment. Mutations observed in these two populations may be associated with different predictors. Hence, it is also a potential caveat of current results.

Identifying predictors of infrequent events is an important but difficult task in studying biological databases. We have addressed this challenge by developing a combined method using association rule analysis to find potentially interesting patterns involving rare events and logistic regression and classification trees to establish linear and high-order associations among those events. In this paper, we show that our approach can identify strong predictors for rare events in a biomedical genomics database and can provide researchers potentially biologically relevant results for further verification.

Acknowledgments

The authors thank Martin O'Connor for his assistance in data preparation. This work was funded in part by a training grant from the National Library of Medicine (5T15LM007033-22).

Reference

1. Rhee, S.Y. et al. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* **31**, 298-303 (2003).
2. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, (Springer-Verlag, 2001).
3. Conklin, D., Jonassen, I., Aasland, R. & Taylor, W.R. Association of nucleotide patterns with gene function classes: application to human 3' untranslated sequences. *Bioinformatics* **18**, 182-9 (2002).
4. Ji, L. & Tan, K.L. Mining gene expression data for positive and negative co-regulated gene clusters. *Bioinformatics* **20**, 2711-8 (2004).
5. Oyama, T., Kitano, K., Satou, K. & Ito, T. Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics* **18**, 705-14 (2002).
6. Tan, P., Steinbach, M. & Kumar, V. *Introduction to Data Mining*, (Addison Wesley, 2006).
7. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* **35**, 352-9 (2002).
8. Quigg, M. et al. Association of antiretroviral resistance genotypes with response to therapy--comparison of three models. *Antivir Ther* **7**, 151-7 (2002).
9. Beerenwinkel, N. et al. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc Natl Acad Sci U S A* **99**, 8271-6 (2002).
10. Cook, N.R., Zee, R.Y. & Ridker, P.M. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med* **23**, 1439-53 (2004).
11. Alexe, G. et al. A Robust Meta-classification Strategy for Cancer Diagnosis from Gene Expression Data. *Proc IEEE Comput Syst Bioinform Conf*, 322-5 (2005).
12. Newman, D.J., Hettich, S., Blake, C.L., & Merz, C.J. (1998). UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. [http://www.ics.uci.edu/~mlearn/MLRepository.html].
13. Liu, B., Hsu, W. & Ma, Y. Integrating classification and association rule mining. in *KDD'98* pp.80-86 (New York, NY, 1998).
14. Li, W., Han, J. & Pei, J. CMAR: Accurate and efficient classification based on multiple class-association rules. in *ICDM'01* pp.369-376 (San Jose, CA, 2001).
15. Chae, Y.M., Ho, S.H., Cho, K.W., Lee, D.H. & Ji, S.H. Data mining approach to policy analysis in a health insurance domain. *Int J Med Inform* **62**, 103-11 (2001).
16. Rhee, S.Y. et al. HIV-1 Protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance. *J Infect Dis* **192**, 456-65 (2005).