Report for 2004 Research Project funded by the ORAU Ralph E. Powe Junior Faculty Enhancement Award Program

Title of the research: **Gene Expression Analysis of HIV-1 Linked p24-specific CD4+ T-Cell Responses for Identifying Genetic Markers.**

Carlotta Domeniconi
Information and Software Engineering Department
George Mason University
carlotta@ise.gmu.edu

Abstract. The Human Immunodeficiency Virus (HIV) presents a complex knot for scientists to unravel. After initial contact and attachment to a cell of the immune system (e.g. lymphocytes, monocytes), the virus causes a cascade of intracellular events. The endproduct of these events is the production of a massive number of new viral particles, death of the infected cells, and ultimate devastation of the immune system. HIV is an epidemic and a crisis in many continents. Since there are many variations of the virus and differences in people's genetic make-up, rapid diagnosis and monitoring of tailored treatments are essential for future medicine. To combat this problem, microarray technology can perform a single scan of thousands of genes. However, without a proper research design and data mining techniques, the results from this technology can be very skewed. Thus, using a normalized, clean dataset (time-series) from the CD4+ T-cell line CEM-CCRF, we designed and implemented hierarchical clustering and pattern-based clustering algorithms to identify specific cellular genes influenced by the HIV-1 viral infection. This research can contribute to the HIV Pharmacogenomics field by confirming HIV genetic markers, which would lead to rapid diagnosis and customized treatments.

Keywords: pattern-based clustering, hierarchical clustering, HIV, gene expression analysis, genetic markers.

Introduction

Since viruses (i.e. human immunodeficiency virus type 1 - HIV-1) can impact a diverse set of host cell's biochemical processes, many of these interactions can be characterized by changes in cellular mRNA levels that could depend on both the stage of infection and the biological stage of the infected cell. For example, viral infection induces the interferon antiviral response, modulates the cell's transcriptional, translational, and trafficking machinery. Thus, the recent emergence of high-density DNA arrays (microarrays and oligonucleotide chips) has revolutionized gene expression studies by providing a means to measure mRNA levels for thousands of genes simultaneously.

In this research we conducted a gene expression analysis, which is a novel approach to identifying and profiling genes related to the pathology and responsiveness of a potential treatment. In the case of HIV-1, where the infection is worldwide and the subtypes are many, measuring the efficacy of a potential treatment in distinct populations from a molecular level is essential. Since people can have different responses to treatments based on their genetic make-up, the Food and Drug Administration is going to mandate pharmacogenomic studies to be submitted with drug submission research.

Thus, we focused on two main objectives:
1. Researching and discussing the various techniques and approaches for gene expression analysis.
2. Identifying and confirming global genetic markers for HIV-1 by designing and implementing data mining algorithms.

Our approach utilized two proven computational techniques: hierarchical clustering and pattern-based clustering. The data analysis is based on time series data and genes from the CD4+ T-cell line CEM-CCRF in order to identify specific cellular genes influenced by HIV-1 viral infection.

Motivation and Contribution

The results of this study can give great insight on how to quickly measure the effectiveness of a treatment according to a person's genetic make-up, and what specific genes are important in the regulation of HIV/AIDS. This study will help to confirm previous results from a molecular level and contribute to the overall knowledge domain of pharmacogenomic-HIV research, which will eventually lead to customized diagnosis and treatment of the disease.

Summary of Results

The human immunodeficiency virus type 1 (HIV-1) infection alters the expression of host cell genes at both the mRNA and protein levels. To obtain a more comprehensive view of the global effects of HIV infection of CD4-positive T-cells at the mRNA level, we analyzed a cDNA microarray dataset generated from the University of California, San Diego. We perform p-clustering and hierarchical clustering analysis on mRNA expressions of approximately 6800 genes. These mRNA expressions were monitored at eight time points [0.5h, 2h, 4h, 8h, 16h, 24h, 48h, 72h] from a CD4+ T-cell line (CEM-GFP) during HIV-1 infection. The CEM-GFP cells were inoculated with HIV-1 at a multiplicity of infection of 0.5, an inoculum sufficient to ensure that every cell is contracted by virus particles. A mock infection served as a control at each time point, essentially replacing the volume of viral input by an equivalent volume of culture medium from uninfected cells. Each sample was tested on two chips and the average was taken. Normalization for this dataset was done using global normalization and scaling. The objective was to identify a specific set of universal genes that can be used as genetic markers for measuring the effectiveness of a potential treatment based on time series patterns and levels consistently changing more than 1.5-fold. A fold is defined mathematically as $\log_2(Cy5/Cy3)$, where typically $Cy5$ represents treated/infected samples and $Cy3$ represents untreated/uninfected samples. Therefore, the expression values are clustered by trends over a period of time and by fold regulation.

From the analysis, we were able to single out six individual genes that could serve as potential genetic markers. The accession number for the first gene is *J04423*. Because this gene was of high interest during the microarray experiment, six different probe sets were used, each resulting in a significant fold regulation by 72 hours. The probe that yielded the highest fold increase had an upfold regulation of 1.85 ($\log_2$ (25448.1/7187.9)) at 72 hours. The next gene - accession number *XO3453* - was analyzed with two different probe sets. The probe that yielded the highest fold regulation had an upfold regulation of 1.55 ($\log_2$ (65440.2/22487.1)) at 72 hours. The other four genes (accession numbers stated below) of interest were only analyzed using one probe set and yielded the following results:

- *U14573*: upfold regulation of 1.5
  ($\log_2$ (95340.6/34555.2)) at 72 hours
- *AB000905*: upfold regulation of 1.5
  ($\log_2$ (210.2.9/76)) at 72 hours
- *D43951*: upfold regulation of 2.45
  ($\log_2$ (111.6/20.7)) at 72 hours
- *M21388*: upfold regulation of 1.5
  ($\log_2$ (28749.2/10162.9)) at 72 hours

  As an example, the plot in Figure 1 shows the expression value for each time point and the overall pattern for all the time points for the gene with accession number *J04423*. The pink line represents infected CEM-GFP cells, while the blue line represents non-infected CEM-GFP cells.
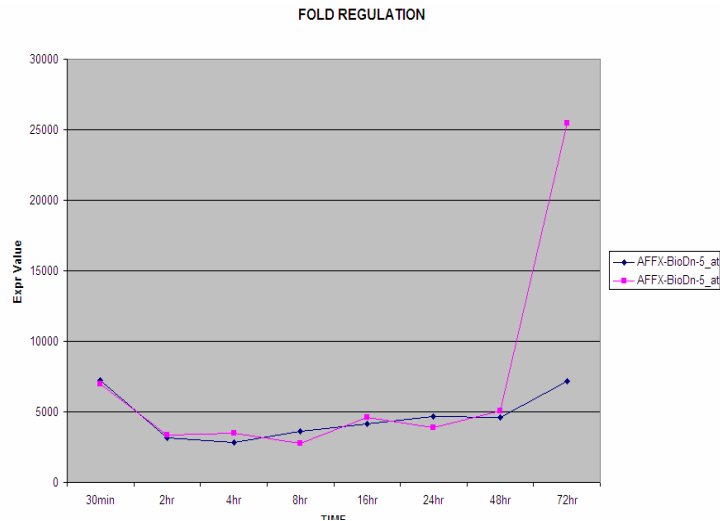
Figure 1: *J04423*

From looking up the six identified genes in the GenBank and NCBI databases, we were able to confirm the results as shown in Table 1.

Accession Number	Gene	Gene Type	Gene Product
J04423	bioD	Protein Coding	Enzyme called dethiobiotin synthetase
X03453	cre	Protein Coding	Enzyme called cyclization recombinase
U14573	Alu	Protein Coding	Actively transcribed by pol III, altered protein sequences
AB000905	HIST1H4I	Protein Coding	Histone 1, H4i
D43951	PUM1	Protein Coding	Assist in RNA binding and mRNA metabolism
M21388	GLA	Protein Coding	Enzyme called alpha-galactosidase

Table 1: Potential genetic HIV-1 markers and their confirmed functionality

Although some of these genes belong to different chromosomes, we can infer that they are affected in a similar fashion when exposed to HIV-1 virus after 3 days. Therefore, one can see why it is important to not only look for co-expressed genes, but also for coherent genes in order to obtain a full snapshot of the gene's profile.

Conclusions

The results obtained are promising, and provide a good starting point for further research in this area. This research can contribute to the HIV Pharmacogenomics field by confirming HIV genetic markers, which would lead to rapid diagnosis and customized treatments. In fact, doctors can easily use these markers, along with other markers for different diseases, to rapidly diagnose a patient's profile in one genetic scan. At the same time, these markers can be used to monitor the progression or treatment of the disease. To further confirm the results obtained, *in-vivo* samples will be used in our future work.

For additional details, see:
S. Raman and C. Domeniconi, "Gene Expression Analysis of HIV-1 Linked p24-specific CD4+ T-Cell Responses for Identifying Genetic Markers", in Proceedings of the ***International Workshop on Feature Selection for Data Mining: Interfacing Machine Learning with Statistics. In conjnction with SIAM International Conference on Data Mining*** , Newport Beach, California, April 21-23, 2005.

Sanjeev Raman is a PhD candidate in Information Technology at George Mason University conducting research on the topics discussed in this report, under my supervision.