

Using Machine Learning to Predict the Health of HIV-Infected Patients

Charles L. Cole
Bucknell University
Lewisburg, PA 17837
charles.cole@bucknell.edu

Brian R. King, Ph.D.*
Bucknell University
Lewisburg, PA 17837
brian.king@bucknell.edu

ABSTRACT

Human immunodeficiency virus-1 is a complex retrovirus that gradually destroys the body's immune system, making it harder for the individual to fight infections. The worst prognosis for an infected individual is AIDS; however this result does not occur with everyone. Moreover, not every infected person develops AIDS at the same rate [5]. We developed a method that can predict the disease prognosis of human HIV infections based on non-redundant HIV genomic and proteomic sequence data. Using the random forest classification method on the genomic data, we obtained over 91% accuracy over four different disease levels. We also analyzed the proteins expressed from five of the nine genes in HIV. We found that the rev gene had the highest predictive performance for disease level. Using a decision tree, we were able to output rules that contained specific variants in the protein that can suggest disease outcomes. This information may help researchers understand underlying variants of the gene that have different patient outcomes. Moreover, this knowledge can improve the selection of appropriate treatment methods depending on the predicted infection level, and also improve drug targeting.

Categories and Subject Descriptors

J.3 [Computer applications]: Life and Medical Sciences—*Biology and Genetics*

Keywords

HIV, sequence analysis, n-gram model, decision tree, random forest

1. METHODS AND MATERIALS

The aim of our research is to determine the relationship between features of the HIV sequence and the progression of disease in the infected person. Genomic sequences were

*Corresponding author.

collected from the Los Alamos National Laboratory HIV database [2]. In total, 76,065 raw sequences were collected. Duplicate sequences and sequences that were missing genomic data were removed, resulting in 46,556 genomic sequences. Sequences were partitioned into four groups based on patient health, namely *acute infection*, *asymptomatic*, *symptomatic*, and *AIDS* (deceased patients were not considered) (Figure 1).

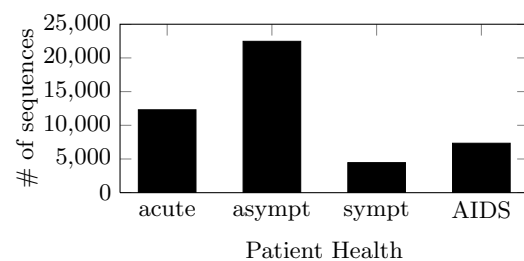


Figure 1: Distribution of genomic sequence data.

1.1 Nucleic Acid

We employed an n -gram model of our sequence data [4], where each sequence is transformed into a frequency occurrence vector over all possible n -grams. In our work, an n -gram is a fixed-length subsequence of length n . The frequency vectors represent our feature space for each sequence. We used Weka [3] with 10-fold cross validation to induce classification models to predict health. We chose the C4.5 decision tree [7] and the random forest [1] classifiers.¹ Table 1 shows the performance of each classifier for different values of n .

¹C4.5 (J48) was run using default settings. Random forest was run using the parameter numTrees=100.

Table 1: Comparison of the effectiveness of various machine-learning algorithms on classifying disease level based on the n -gram model of the genome. The overall accuracy and weighted average F_1 performance metrics are displayed.

n	C4.5		Random Forest	
	Acc	F_1	Acc	F_1
1	83.39%	0.832	86.0577%	0.859
2	85.99%	0.859	90.3084%	0.902
3	86.47%	0.864	91.2879%	0.911
4	86.33%	0.863	91.5457%	0.914
5	86.33%	0.863	91.8249%	0.917

The random forest method had the best performance, with a peak correct classification rate of 91.83%, and a false positive rate of only 0.008 in symptomatic health classification. Higher values of n did not achieve superior results in our tests.

1.2 Proteins

We wanted to delve into protein sequence data to determine if specific gene products had stronger correlations with disease levels. We looked at five of the nine genes in HIV, including *vif*, *vpr*, and *vpu* proteins, which are nonstructural genes that are unique to HIV, *rev*, which is required for nuclear export, and *tat*, which increases the level of transcription of HIV [6]. Only 8% of the genomic data download had corresponding protein data available, yielding an average of 832 sequences per gene. See poster for dataset details.

For this experiment, we adopted a different model of the protein sequence, where we partitioned each sequence into k equal-length partitions. Each partition overlapped with adjacent partitions by three positions in the sequence, to compensate for minor variation observed in sequence lengths. The amino acid composition was computed for each partition (equivalent to a 1-gram model). The vectors for each partition were concatenated and used to represent each protein for classification. For our tests, we varied k from 1 to 10. Results are shown in (Figure 2).

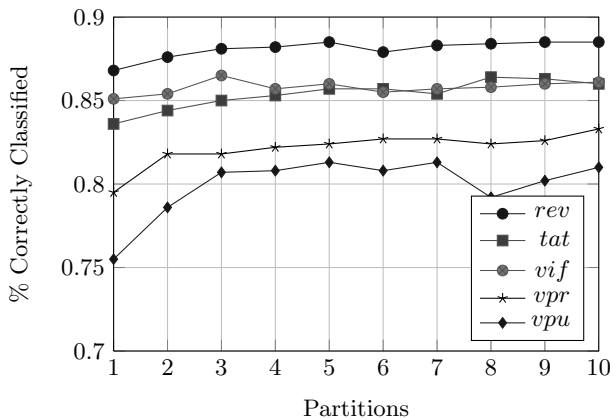


Figure 2: Classification of sequences classified using a random forest. All Partitions had an overlap of zero with one another.

We observed that the *rev* gene had the best performance, with an overall accuracy ranging from 86.8% - 88.5%. This

could be due to the fact that *rev* plays an important role in exporting intron-containing HIV RNA from the nucleus [6], so little variation among the gene is possible in order for it to function correctly. An important reason we chose to partition the sequence is to develop a method that can identify regions of the sequence that have the strongest predictive power of disease phenotype. By applying a tree-based classifier, such as C4.5, a visual representation of the decision tree is able to be created (Figure 3). A representation such as this may help AIDS researchers by allowing them to observe strong amino acid propensities in specific regions of the protein sequence for a specific gene that are correlated with a specific disease phenotype.

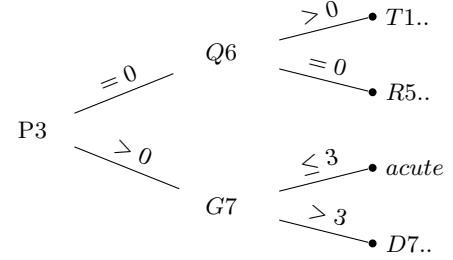


Figure 3: This figure shows a small section of a decision tree used to classify the sequences. The letter indicates the amino acid and the number following it refers to the partition. Based on that count, the tree is traversed until the sequence is classified.

2. ACKNOWLEDGMENTS

Part of this research was supported through the Program for Undergraduate Research at Bucknell University.

3. REFERENCES

- [1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [2] K. C. F. B. L. T. A. C. H. B. M. I. M. J. R. A. W. S. and K. B. Hiv sequence compendium. *Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR 10-03684*, 2010.
- [3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.
- [4] B. King and C. Guda. ngloc: an n-gram-based bayesian method for estimating the subcellular proteomes of eukaryotes. *Genome Biology*, 8(5):R68, 2007.
- [5] M. Koot, I. P. M. Keet, A. H. V. Vos, R. E. Y. de Goede, M. T. L. Roos, R. A. Coutinho, F. Miedema, P. T. A. Schellekens, and M. Tersmette. Prognostic value of hiv-1 syncytium-inducing phenotype for rate of cd4+ cell depletion and progression to aids. *Annals of Internal Medicine*, 118(9):681–688, 1993.
- [6] V. W. Pollard and M. H. Malim. The hiv-1 rev protein. *Annual Review of Microbiology*, 52(1):491–532, 1998. PMID: 9891806.
- [7] J. R. Quinlan. *C4.5: Programs for Machine Learning*, volume 1. Morgan Kaufmann, 1993.