Comparison of HTLV and STLV by Using Apriori Algorithm and Decision Tree

Jinwon Kwon, Subin Yoon, Cheryn Kim, Sang Ryul Kim, and Taeseon Yoon
Hankuk Academy of Foregin Studies, Mohyen-myeon, Yongin-si, Republic of Korea
Email: paige1008@naver.com

***Abstract*—In this study, Apriori algorithm and Decision tree C5.0 algorithm were employed to compare the sequence patterns of HTLV(Human T-lymphotropic) and STLV(Simian T-lymphotropic) viruses. HTLV and STLV are Deltaretroviruses and are counterparts. However, these viruses have a complex history. One used to be HIV and some were found and named recently. To find the congruity of these viruses, we decided to compare them. Amino acids of the HTLVs showed that pattern of HTLV-1 and 2 could also be applied to HTLV-3, but HTLV-4 has a different pattern. Rules of the sequences tend to follow HTLV-2's rule, yet it is too subtle to come into consideration.**

***Index Terms*—HTLV(Human T-lymphotropic), STLV(Simian T-lymphotorpic), apriori, decision tree, deltaretroviruses, compare**

I. Introduction

Deltaretroviruses are a group comprised of Bovine leukemia virus (BLV), human T-lymphotropic viruses (HTLV), and simian T-lymphotripic viruses (STLV). HTLV has 4 strains; HTLV-1, HTLV-2, HTLV-3 and HTLV-4. HTLV-1 and 2 with their simian counterparts STLV-1, 2 and 3 fit in the Primate T lymphotropic viruses group (PTLV) [1]. HTLV-1 and STLV-1 are highly related; HTLV-2 and HTLV-3 are closely related to STLV-2 and STLV-3, respectively. Thus, the HTLVs are considered to have been develped from interspecies transmission between monkeys and humans. The genetic diversity of HTLV-1 strains is less than 8%, and HTLV-1 and HTLV-2 show 70% nucleotide homology [2]. The discovery of HTLV-1 was in the 1980s when retrovirus particles were found in T-cell lymphoblastoid cell lines and in fresh peripheral blood lymphocytes attained from a subject with T cell lymphoma and also from adult T cell leukemia patient [3]. The unique protein expressed by ribonucleic acid in HTLV-1 is potentially carcinogenic. Subsequently, HTLV-2 was found in a patient with a T-cell variant of hairy cell leukemia [5], [6]. Adult T cell leukemia (ATL) and myelopathy/tropical spastic paraparesis (HAM/TSP) are known as HTLV-1 linked disease. To this date, no related disease has been revealed in the case of HTLV-2, HTLV-3, and HTLV-4. In rare cases, however, HTLV-2 shows some association with neurological disorders [3]. At present there is no treatment to eradicate the infection. Retroviral genome contains codes for producing proteins necessary for building up a virus. Gag proteins for capsid, nucleocapsid and matrix, Pro proteins, polymerase proteins (Pol) and Env proteins are major proteins and ard make from complete of once spliced mRNAs. Other regulatory and accessory proteins come from alternatively spliced mRNAs. There are two regulatory genes called tax and rex; tax is a transactivation gene responsible for transforming primary human T cells in HTLV1 and 2. and rex is post-transcriptional gene by blinding and stabilizing intron containing mRNAs. Among four types of HTLV, HTLV-1 and 2 are the most prevailing viruses around the world. Other types of human lymphotropic viruses have a relatively shorter history. HTLV-3 and 4 have recently been identified from South of Cameroon in 2005 [7]. Accordingly, HTLV-3 and 4 are less researched than HTLV-1 and 2 such as the prognosis or treatments of the infection. By analyzing PTLVs of differences and similarities, the connection between HTLV-1, 2 and HTLV-3, 4 can be made more understandable. This should provide insights into better prognosis and prevention strategies in the yet-discovered field of PTLVs.

II. Methods

A. Apriori Algorithm

Apriori algorithm is usually applied in mining process and association rule finding of transactional data. It sorts out repeated items in a section and search for more repetition of the same item by going further on a larger part of the database. Through the process, the general trend of the data can be seen vividly. It also enables comparing association rules among several data sets. It is one of the algorithms that has appropriate match with the deoxyribonucleic acid (DNA) sequence or RNA sequences because such sequences are linear and does not have timelapses. They also have distict repeated code patterns that later is translated into amino acids. To use the apriori algorithm, we first extracted exons; exon is the meaningful area of the DNA sequence that contains the actual genetic information which is later transcribed in order to make protein products. We compared the number of base sequences in deltaretroviruses that are used to code for various types of amino acids. Here, we used 9, 13, and 17-window to search for amino acids that appear often in each window. Window deviding system breaks sequences into small parts containing a number of sequences that is identified by the number of the window.

Manuscript received May 30, 2015; revised July 28, 2015.

doi: 10.18178/ijpmbs.4.3.163-166

If window 13 is used, 13 bases are included in one fragment.

B. Decision Tree

Decision tree is a rule mining algorithm which is used to support classification process. It is frequently used to classify numerous data which include numerical value variable or categorical variable. It categorizes data from making root nodes. Root nodes are the most significant node that doesn't have a parent node. Starting from the root nodes, branches are made in order to make a decision (with binary code) from parent node connect another nodes. This process is repeated until it reaches the final nodes (left node). The final node results are printed as the results. A decision is clear, and show well-ordered list, so it is easy to understand the data structure. Fig. 1 shows a very simplified version of a decision tree, which decides whether the people will play or not according to the weather.

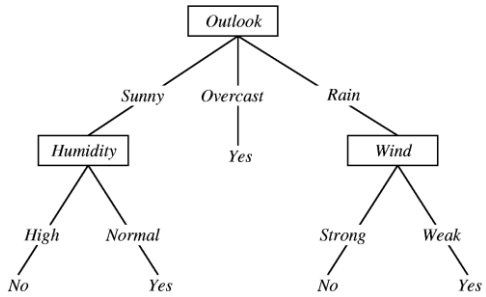


Figure 1. Simple decision tree.

Decision tree algorithm is a rule mining algorithm which is frequently used to datamine the DNA sequences. In this study, we used 10 fold cross-validation method and rule-based classifiers. 10 fold cross-validation method is a method where one takes one of the 10 datasets to extract rules from it, and then applies the rules to other 9 sets to get the percentage of synchronization as the result. This is why some rules are repeated constantly in several folds. Then, for each class, we distilled rule sets by drawing out rules whose frequency values are over 0.8. We obtained complete sequences of all the PTLVs from NCBI to compare the nucleic patterns. In the case of HTLV-, however, we obtained partial DNA sequence of HTLV-4, as no complete sequence was provided.

III. Results and Discussion

A. Apriori Results

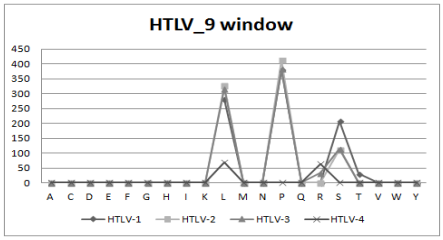


Figure 2. 9 window HTLV.

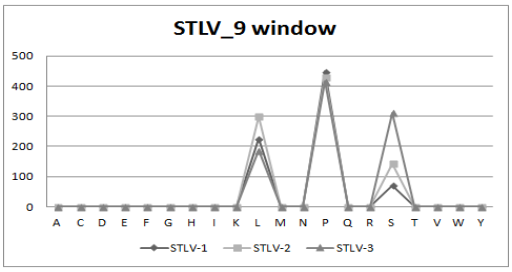


Figure 3. 9 window STLV.

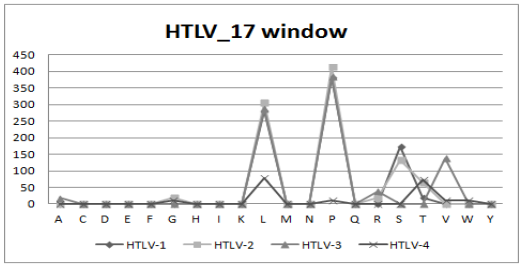


Figure 4. 13 window HTLV.

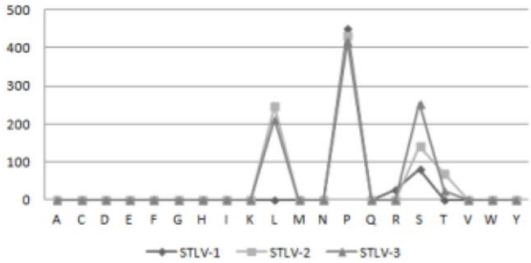


Figure 5. 13 window STLV.

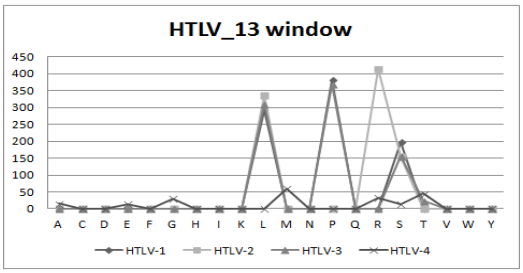


Figure 6. 17 window HTLV.

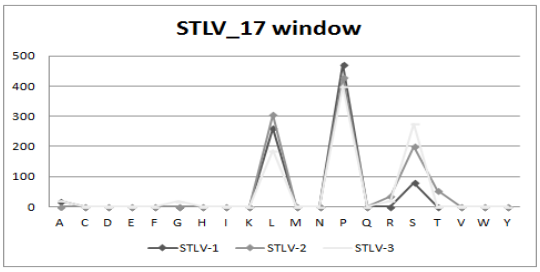


Figure 7. 17 window STLV.

In a single type analysis, some analysis can be made on HTLVs and STLVs based on their window results. HTLV, According to Fig. 1, 2 and 3, the, HTLV-1,2, and 3 shows analogous values on Leucine and Proline. Serine is also often observed. Also, Arginine and Valine is observed in some considerable amount in types of viruses. When it comes to HTLV-4, however, its tendency distinguished

clearly from others [8]. This pattern is quite distinct and even the number of proteins are very alike. Thus, the graph shows comparable habit of HTLV-1, 2 and 3 and HTLV-4 with comparatively distinctive pattern STLV, Fig. 4 and 5 shows the samples highly mathching, but when we take a look at Fig.6, it is quite surprising how those graphs are matching almost identically with each other, as if they were one in the first place. They all have same three proteins with no extra proteins; they have Leucine, Proline and Serine. Other windonevertheless similar. There are almost no extra proteins produced. Just like in HTLVs, we can also see the amounts of proteins are similar in STLV graphs.

Comparing the two types, HTLV3 and STLV3, Going through the graphs, we found out that HTLV3 and STLV3 had a quite similar pattern. So we tried combining the two graphs. As a result, HTLV3_13window and STLV3_17window showed an astonishing match. Types of protein and even the number was almost perfectly fitting with another.

Comparing the two types, HTLV3 and STLV3, Going through the graphs, we found out that HTLV3 and STLV3 had a quite similar pattern. So we tried combining the two graphs. As a result, HTLV3_13window and STLV3_17window showed an astonishing match. Types of protein and even the number was almost perfectly fitting with another.

B. Decision Tree Results

This experiment used by See 5.0 program.

Each class represents different types of viruses of the given type. For example, class 2 of HTLV group represents HTLV-2 virus.

C. Discussion

The data shown in numbers is the number of rules that the classes share. The more rules the classes share, the more similar the classes are. In 9 window STLV results, the number of rules shared is quite concentrated, showing a number between 94 and 131. All of them are highly similar. In 13 window STLV results, the results are once again concentrated around 80 and shows high similarity since 80 rules shared is still a very considerable number of rules. 17 window STLV resulted in 41 to 83 rules, however 17 bases for one group of data is a large scale, and it is natural to expect relatively smaller number of shared rules and a wider span. Overall, there is no tend in the data. The classes are equally similar to each other.

In 9 window HTLV results, class 1 to class 3 show a very high number of rules shared. The number appears between 72 to 112, which is very concentrated and high, showing that these classes are both very similar and equal. No one class can be a bigger category for another. In contrast, class 4 does not follow other classes nor is followed by them. 35 to 59 rules shared is not that low a number, however compared to the similarity of class 1 to 3, it can be considered subtle. 13 window HTLV shows a smaller gap between coherency of classes 1 to 3 and coherency with class 4, where results for classes 1 to 3 starts from 45 rules and the highest result for class 4 is 46. But when taking a mean of each catgory and then comparing, class 4 is still very isolated. Like STLV, higher window number, which represents bigger cut data to work with, resulted in smaller number of rules shared. Somehow, results in 17 window HTLV showed surprisingly high number of rules shared, some even showing a number higher than 100. But class 4 still shows a low result that averages in 35 rules shared. The abnormally high result in 17 window must indicate high similarity in the HTLV viruses.

The viruses tended to follow the rule of HTLV2, but the pattern is so subtle that the result is meaningless. Table I to Table VII all show a fair amount of rules following each other. Therefore, the viruses can be seen as not belonging to any one virus type.

TABLE I. AMINO ACID CHART

Amino Acid	Letter
Alanine	A
Cysterine	C
Aspartic acid	D
Glutamic acid	E
Phenylalanine	F
Glycine	G
Histidine	H
Arginine	R
Serine	S
Threonine	T
Selenocysteine	U
Isoleucine	I
Lysine	K
Leucine	L
Methionine	M
Asparagine	N
Proline	P
Glutamine	Q
Valine	V
Tryptophan	W
Tyrosine	Y

TABLE II. 9 WINDOW STLV

classified	Class1	Class2	Class3
Class1	131	94	110
Class2	115	118	95
Class3	128	97	106

TABLE III. 9 WINDOW HTLV

classified	Class1	Class2	Class3	Class4
Class1	79	108	89	35
Class2	86	114	90	42
Class3	112	103	72	44
Class4	49	59	50	42

IV. CONCLUSION

Overall, our study focused on amino acid patterns in DNA sequences of PTLVs. The results show that HTLV1 and 2's traits, which have been studied relatively longer

than other types of HTLVs, can be adapted to relatively recently found HTLV 3. With this linkage found, we hope it would contribute to gaining insights in the ongoing prognosis and therapeutic agent development. Only, in the case of HTLV-4, we observed a distant nucleic pattern from other types of HTLVs. Thus it would be more plausible for HTLV-4 to be researched directly on a more independent project. Furthermore, there is a possibility to classify HTLV-4 as another group other than the PTLV family. Also, it is considerable to use not the complete sequence of the viruses, but certain sequences such as Env of Gag to reach meaningful results on the specific functions and comparison to genes of other viruses. By analyzing PTLVs of differences and similarities, the connection between HTLV-1, 2 and HTLV-3, 4 can be made more understandable. This should provide insights into better prognosis and prevention strategies in the yet-discovered field of PTLVs.

TABLE IV. 13 WINDOW STLV

classified	Class1	Class2	Class3
Class1	69	92	71
Class2	74	83	70
Class3	77	83	69

TABLE V. 13 WINDOW HTLV

classified	Class1	Class2	Class3	Class4
Class1	47	93	45	30
Class2	55	80	64	31
Class3	57	95	46	31
Class4	25	46	34	34

TABLE VI. 17 WINDOW STLV

classified	Class1	Class2	Class3
Class1	68	60	49
Class2	83	49	42
Class3	82	41	52

TABLE VII. 17 WINDOW HTLV

classified	Class1	Class2	Class3	Class4
Class1	79	108	89	35
Class2	86	114	90	42
Class3	112	103	72	44
Class4	49	59	50	42

REFERENCES

[1] J. Coffin, A. Haase, J. A. Levy, L. Montagnier, S. Oroszlan, *et al.*, "What to call the AIDS virus?" *Nature*, vol. 321, no. 6065, p. 10, May 1986.

[2] W. M. Switzer, M. Salemi, S. H. Qari, H. Jia, R. R. Gray, *et al.*, "Ancient, independent evolution and distinct molecular features of the novel human T-lymphotropic virus type 4," *Retrovirology*, vol. 6, no. 1, p. 253, 2009.

[3] B. J. Poiesz, F. W. Ruscetti, A. F. Gazdar, P. A. Bunn, J. D. Minna, and R. C. Gallo, "Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma," *Proc Natl Acad Sci*, U. S. A., vol. 77, pp. 7415-7419, 1980.

[4] P. Kannian and P. L. Green, "Human T Lymphotropic Virus type 1 (HTLV-1): Molecular biology and oncogenesis," *Viruses*, vol. 2, no. 9, pp. 2037-2077, Sep 2010.

[5] V. S. Kalyanaraman, M. G. Sarngadharan, M. Robert-Guroff, I. Miyoshi, D. Golde, and R. C. Gallo, "A new subtype of human T-cell leukemia virus (HTLV-II) associated with a T-cell variant of hairy cell leukemia," *Science*, vol. 218, pp. 571-573, 1982.

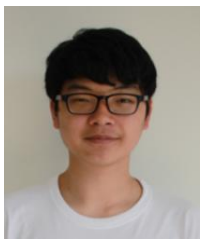
[6] B. Hjelle, O. Appenzeller, R. Mills, S. Alexander, N. Torrez-Martinez, R. Jahnke, and G. Ross, "Chronic neurodegenerative disease associated with HTLV-II infection," *Lancet*, vol. 339, pp. 645-646, 1992.

[7] S. Calattini, S. A. Chevalier, R. Duprez, S. Bassot, A. Froment, *et al.*, "Discovery of a new human T-cell lymphotropic virus (HTLV-3) in central Africa," *Retrovirology*, vol. 2, p. 30, 2005.

Jinwon Kwon was born in South Korea in 1997. She is currently a student of Hankuk Academy of Foreign Studies, Republic of Korea. She is dedicated with biology and computer science, and has interest in immunology recently. She has written studies about viruses using various algorithms previously, using their base pair sequences.

Subin Yoon was born in South Korea in 1997. She is currently a student of natural science program in Hankuk Academy of Foreign Studies, Republic of Korea. She has written studies on viruses using various algorithms previously, using their base pair sequences.

Cheryn Kim was born in South Korea in 1997 and is currently a student at Hankuk Academy of Foreign Studies, Republic of Korea. She has most recently written a paper regarding analyzation of virus genome using particular computer algorithms. She is interested in Paleopolyploidy.



Sang Ryul Kim was born in Seoul, Korea, in 1997 and is currently a student of student of natural science program in Hankuk Academy of Foreign Studies, Republic of Korea. He is interested in computer science and engineering, especially in system trading, bioinformatics, information security.

Taeseon Yoon was born in Seoul, Korea, in 1972. He was Ph.D. Candidate degree in Computer education from the Korea University, Seoul, Korea, in 2003.

From 1998 to 2003, he was with EJB analyst and SCJP. From 2003 to 2004, he joined the Department of Computer Education, University of Korea. Since December 2004, he has been with the Hankuk Academy of Foreign Studies. He is mostly interested in Bioinformatics.