Analysis of Amino Acid Sequence Patterns of Retroviruses

Yoondeok Jeon, Jiwoo Oh, Yongha Jo, and Taeseon Yoon
Hankuk Academy of Foreign Studies, Wangsan-ri, Mohyun-myeon, Yongin-si, Gyeonggi-do, South Korea
Email: {junyd5469, ojw0414, yonghajo}@gmail.com, tsyoon@hafs.hs.kr

***Abstract*—Human Retrovirus, including Human Immunodeficiency Virus(HIV) that causes an infection which develops to a disease of human immune system, affects a lot of kinds of species, including fish, horse, mammals, and even humans. In this paper, we analyzed the amino acid sequence of 5 retroviruses: HIV-1, Human T-lymphotropic virus-1 (HTLV-1), Avian Leukosis virus (ALV), Rous sarcoma virus (RSV), and Equine infectious anemia virus (EIAV), and investigated the difference between them based on the genus they are involved in (HIV-1 and EIAV: lentivirus, ALV and RSV: alpharetrovirus, HTLV-1: deltaretrovirus). Furthermore, we compared the amino acid sequences of HIV-1 and HTLV-1, the most well-known retroviruses for infecting humans.**

***Index Terms*—retrovirus, alpharetrovirus, deltaretrovirus, lentivirus, HIV-1, HTLV-1, ALV, RSV, EIAV, apriori algorithm**

I. Introduction

Human Immunodeficiency Virus (HIV) causes an infection which develops to a disease of human immune system. A person infected to this virus may experience various kinds of infectious diseases and tumorigenesis. This disease is called Acquired Immunodeficiency Syndrome (AIDS) [1], [2]. Once infected, the number of CD4 positive T lymphoma is hugely decreased which causes an enormous decrease in immunity [3]. Since its discovery, overall 34 million cases have been reported, and as of 2009, it has caused nearly 30 million deaths globally. Nowadays, the red ribbon presented in Fig. 1 represents AIDS.

Figure 1. Red ribbon which represents AIDS

HIV/AIDS has a great impact on societies, as an incurable disease and a source of discrimination. Furthermore, since its identification in 1980s, HIV/AIDS has significantly attracted medical and political attentions internationally, with large-scale funding, yet no cure was made.

HIV belongs to the Retroviridae, a family of RNA viruses using reverse transcriptase (RT) for their replication [4]. Fig. 2 shows the life cycle of a retrovirus. Retroviruses replicate their single-stranded RNA as a template with RT to double-stranded DNA, which is a reverse process to original, thus "retro."

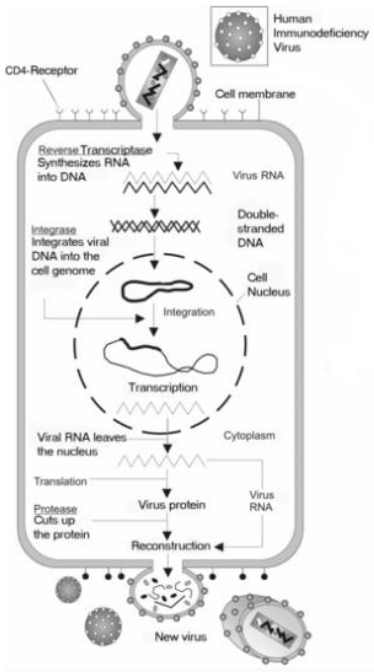

Figure 2. Life cycle of HIV

As HIV/AIDS shows, retroviruses are unstable and cause frequent mutations [5]. This means developing a cure for their infections is almost impossible. We, therefore decided to study about the whole family retroviridae, including HIV, Human T-lymphotropic virus (HTLV), Avian Leukosis virus (ALV), Rous sarcoma virus (RSV), and Equine infectious anemia virus (EIAV) by analyzing amino acid sequences. Due to their functional and structural similarities, we assumed that they might show similarities in their amino sequences, too.

II. Materials

Retroviruses are divided into six genera: Alpharetrovirus, Betaretrovirus, Gammaretrovirus,

Manuscript received July 1, 2012; revised October 28, 2014.

doi: 10.12720/jomb.4.5.426-429

Epsilonretrovirus, Deltaretrovirus, Lentivirus, and Spumavirus. Our study focuses on analyzing differences and common factors from various DNA sequences of infectious retroviruses: HIV-1 (Lentivirus), HTLV-1 (Deltaretrovirus), ALV (Alpharetrovirus), RSV (Alpharetrovirus), EIAV (Lentivirus)

Lentiviruses are featured by long incubation period, 2 years for HIV as an example. These viruses have the ability of carrying a substantial amount of viral RNA into the DNA of host cells. Unlike other retroviruses, lentivirus attacks non-dividing cells [6]. One of the incurable diseases brought by HIV is AIDS. With the two main characteristics of lentivirus, after the incubation period, it starts budding out of cell membrane. Patients die from AIDS because of weak immune system and complication of other illnesses. Unlike other retroviruses, HIV is slightly different from those in structure. Little bigger than normal viruses, it includes two copies of positive single-stranded RNA which is coded for its nine genes enveloped by a conical capsid consisted of thousands of replications of the viral protein p24 [7]. Equine infectious anemia Virus (EIAV), another member of lentivirus, has the smallest and the genetically simplest lentiviral genome. It can be transmitted via blood, milk, and body secretions, and interestingly, EIAV-infected animals develop highly effective immune response and control viral replication and disease. It suggests that effective vaccines against EIAV can be made, and the vaccine developed in china is currently being widely used [8].

Alpharetroviruses, which have type C morphology, are known to penetrate into cells of wild and domestic birds and cause sarcomas, tumors, or anemias. They include ALV and RSV. Rous sarcoma virus (RSV) is one of retroviridae that causes sarcoma in chickens. It is the first oncovirus in the history of virus discoveries; [9] the virus extract was found to induce oncogenesis. It is classified as an alpharetrovirus with type C morphology. Avian Leukosis virus (ALV) is genetically closely connected to Rous sarcoma virus (RSV). Both contain the common factors of retroviruses, the gag gene which encodes for the capsid proteins and the pol gene which encodes for the reverse transcriptase enzyme. ALV is also the one that replicates in chicken embryo fibroblasts but does not transform it compared to RSV [10].

Deltaretroviruses consist of exogenous horizontally-transmitted viruses found in few types of mammals. One of the well-known deltaretroviruses is Human T-Lymphotrophic virus type 1 (HTLV-1) that is known to be an exclusive causal agent of adult T-cell leukaemia [11]. It includes rex, p21, Tax protein, and Tax is known to play a major role in HTLV-1 infection and transmission, cell survival, multipolar mitosis, and aneuploidy. It has three transmission methods: mother-infant, sexual contact, and parental transmission [12]. There are several other types of HTLVs, including HTLV-3 and HTLV-4 [13], and HTLV-2 is known to share 70% genomic homology with HTLV-1.

Currently, the classification of retroviruses above is done in the basis of morphological differences. We analyzed the features of 3 genera of retroviruses and 5 types of retroviruses respectively, and found some similarities between each genus of retroviruses: lentivirus, alpharetrovirus, and deltaretrovirus. Furthermore, it is well known that different viruses infect different species as their hosts. Thus, we've made two hypotheses: First, each genus of retroviruses will share some similarities in amino acid sequences, and second, viruses that share same species as their host will also share similarities in their amino sequences, too. To prove the hypotheses, we analyzed the amino acid sequences of HIV-1, EIAV (lentivirus), RSV, and ALV (alpharetrovirus). Additionally, we've compared the amino acid sequences of HIV-1 and HTLV-1 since they are the two retroviruses best known for infecting human species.

III. Experiment

Data: Five retroviruses whose genomic sequences are available were downloaded from the Genbank (NCBI, http://www.ncbi.nlm.nih.gov). They are NC_001802 (HIV-1), NC_001436 (HTLV-1), NC_015116 (ALV), NC_001407 (RSV), and NC_001450 (EIAV)

Apriori Algorithm: Our analysis bases on Apriori Algorithm, suggesting common factors and major differences between the five retroviruses. Apriori Algorithm is known as a classic form of algorithm for finding and extending frequent items in a database. We could highlight the general trends of the database and understand the association rules [14].

Most algorithms that predict the secondary structure of proteins, including Apriori algorithm we used in our study, cut the amino acid sequence into parts with certain window size, and predict the secondary structure at the center of the window by the features of each amino acid sequence. We converted the genomic sequencesof viruses into amino acid sequences, and divided the sequences into 5, 7, 9 windows respectively and analyzed the sequence pattern and the correlation with the secondary structure that is far from the center of the window at a distance of 'd.' The results of datamining is in the below. For example, the results showed these kinds of patterns.

TABLE I. BEST RULES FOUND FOR HIV-1 IN 9WINDOW

HIV-1
1. Amino2 is L 49
2. Amino6 is L 46
3. Amino7 is P 45
4. Amino8 is P 44
5. Amino5 is P 43
6. Amino5 is L 42
7. Amino2 is P 41
8. Amino3 is P 41
9. Amino3 is S 41
10. Amino4 is P 40

These are parts of the result shown by analyzing HIV-1 retrovirus divided into 9 windows. 100 rules were found by each retrovirus, so we got 1500 rules in our experiment. (5 retroviruses each were divided into 5, 7, 9 windows respectively) To analyze the rules, we've listed all the rules based on their frequencies, and the example

is on the below (Table II). Next, we selected two amino acid lines randomly, and compared the viruses with same amino acid line. For example, we selected amino1 and amino2 for analyzing HIV-1 and EIAV as lentivirus family, so we compared amino1 and amino2 of each viruses (amino 1 and 2 of HIV-1 and EIAV).

TABLE II. LIST OF AMINO ACIDS FOR EACH AMINO ACID LINE IN EIAV

EIAV
amino1={I,G,N,L,R,S,K,V,Y,T,D}
amino2={R,L,N,S,G,K,V,I,E,T,Y,Q,C,R}
amino3={R,S,L,N,T,I,Y,G,K,V,A,C}
amino4={S,L,G,K,T,E,V,Y,I,N,}
amino5={L,R,K,I,T,S,E,A,G,N,V,W,Y,F,H}
amino6={T,R,K,G,L,I,N,V}
amino7={R,S,L,K,N,T,G,S}
amino8={R,L,T,G,K,I,S,N,H}
amino9={K,L,N}

Based on the results above, for each retroviruses, we arrayed the types of amino acids appearing in each windows of retrovirus based on frequencies of appearance. The results are in the below.

The result suggests that ALV and RSV, two major retroviruses that are involved in alpharetrovirus family, show similarities in the types of amino acids that consists each retrovirus and how they are arranged. From amino 1 to amino 9, two retroviruses share similarities in most windows (the graph shows 2 amino acids of them) when they are listed in frequency-descending order: for most windows, the frequency of appearance for G, A, R, L amino acids was high above others. Furthermore, this tendency was also found in lentivirus family, HIV-1 and EIAV. The graph looks similar in most parts, for most amino acid arrays.

However, the graph of alpharetrovirus and lentivirus differs in many ways. The type of amino acids that are comprised in each family differed, and the frequency of the amino acids appearing in the array didn't show much similarities. The amino acids appearing in the array didn't show much similarities. (The variables that comprise the x-axis are identical.

In addition, we compared the results of HIV-1 and HTLV-1 sequences, which infect humans. As shown, the comparison between them showed some similarities when they were divided into 5window and then analyzed, but when they were divided into 9 windows, they didn't show significant similarities. The types of amino acids composing them were different in many portions of them, and the frequency of appearance was also different. We thought that retroviruses that have same host animals will share similarities in their amino acid sequence, but the result shows that the sequence is not much affected by this criterion.

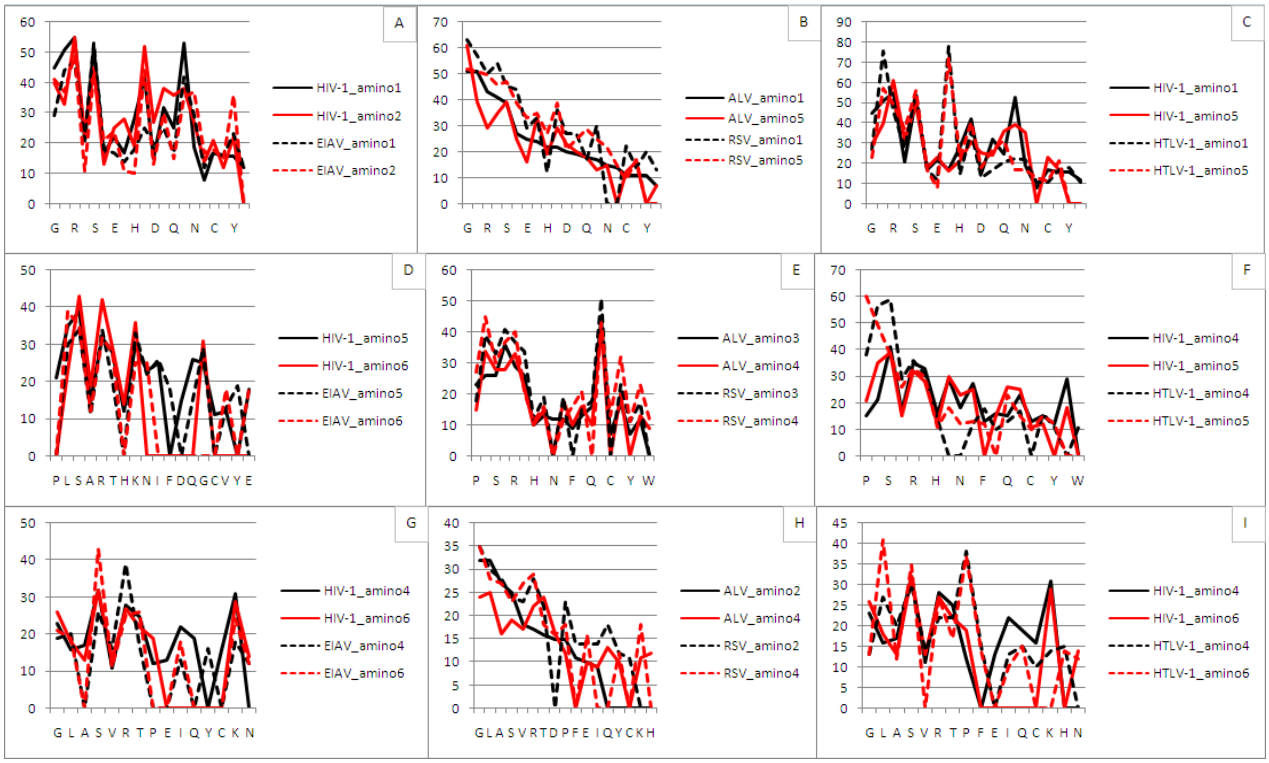

Figure 3. Analysis of 5 retrovirus amino acid sequences. A,B,C: 5 windows, D,E,F: 7 windows, G,H,I: 9 windows. A: comparison between HIV-1 amino1, HIV-1 amino2, EIAV amino1, EIAV amino2. B: comparison between ALV amino1, ALV amino5, RSV amino1, RSV amino5. C: comparison between HIV-1 amino1, HIV-1 amino5, HTLV-1 amino1, HTLV-1 amino5. D: comparison between HIV-1 amino5, HIV-1 amino6, EIAV amino5, EIAV amino6. F: comparison between ALV amino3, ALV amino4, RSV amino3, RSV amino4. F: comparison between HIV-1 amino4, HIV-1 amino5, HTLV-1 amino4, HTLV-1 amino5. G: comparison between HIV-1 amino4, HIV-1 amino6, EIAV amino4, EIAV amino6. H: comparison between ALV amino2, ALV amino4, RSV amino2, RSV amino4. I: comparison between HIV-1 amino4, HIV-1 mino6, HTLV-1 amino4, HTLV-1 amino6.

IV. Conclusion

To date, retroviruses are classified based on their morphological differences, but as a global trend that argues the analysis of amino acid sequence patterns as a classification criteria, which is also called "molecular biological proof", there should be a shift. Thus, we've analyzed 5 different amino acid sequences of retroviruses, HIV-1, HTLV-1, EIAV, ALV, and RSV and searched for similarities.

Before the analysis, we've set two hypotheses: First, each genus of retroviruses will share some similarities in amino acid sequences, and second, viruses that share same species as their host will also share similarities in their amino sequences. According to the result, we found that each genus of retrovirus share similarities in amino acid sequences. The results verify this: two alpharetroviruses and two lentiviruses each share similarities in their amino acid composition, but there are many differences between alpharetroviruses family and lentivirus family in their amino acid sequence patterns. This means that the comparison of amino acid sequence patterns of retroviruses can be a new classification criterion for retroviruses too, as they are already being used in other fields of taxonomy.

Furthermore, considering the comparison between HIV-1 and HTLV-1, we can infer that the type of animal they can use as a host animal may not influence the type of amino acids that comprises the virus nor the amino acid sequences. Even they showed some similarities when they are divided into 5window, it is known that the size of a window might influence the accuracy of the analysis. In other words, the bigger the window, the better the accuracy. Thus, considering the fact that the amino acid sequences of HIV-1 and HTLV-1 didn't show much similarities when they are divided into 9window, it can be concluded that they cannot be classified by similarities among amino acid sequence patterns. These results strongly suggest that by analyzing all types of retroviruses, we can discover the differences between each genus in their amino acid sequences, and this can be used as a new criterion for classifying retroviruses. This would be the ultimate goal of successive researches.

References

[1] R. A. Weiss, "How does HIV cause AIDS?" *Science*, vol. 260, no. 5112, pp. 1273–1279, May 1993.

[2] D. C. Douek, M. Roederer, and R. A. Koup, "Emerging concepts in the immunopathogenesis of AIDS," *Annu. Rev. Med.*, vol. 60, pp. 471–84, 2009.

[3] A. Cunningham, H. Donaghy, A. Harman, M. Kim, S. Turville, "Manipulation of dendritic cell function by viruses," *Current Opinion in Microbiology*, vol. 13, no. 4, pp. 524–529, 2010.

[4] Y. H. Zheng, N. Lovsin, and B. M. Peterlin, "Newly identified host factors modulate HIV replication," *Immunol. Lett.*, vol. 97, no. 2, pp. 225–34, 2005.

[5] D. L. Robertson, B. H. Hahn, and P. M. Sharp, "Recombination in AIDS viruses," *J. Mol Evol.*, vol. 40, no. 3, pp. 249–59, 1995.

[6] J. A. Lee, J. A. Conejero, J. M. Mason, *et al*., "Lentiviral transfection with the PDGF-B gene improves diabetic wound healing," *Plast. Reconstr. Surg*., vol. 116, no. 2, pp. 532–538, August 2005.

[7] Various (2008). HIV Sequence Compendium 2008 Introduction. Retrieved March 31, 2009.

[8] C. Leroux, J. L. Cadoré, and R. C. Montelaro, "Equine Infectious Anemia Virus (EIAV): what has HIV's country cousin got to tell us?" *Vet. Res*., vol. 35, pp. 485-512, 2004.

[9] R. A. Weiss, P. K. Vogt, "100 years of Rous sarcoma virus," *J. Exp. Med*., vol. 208, no. 12, pp. 2351–2355, November 2011.

[10] Weiss, Robin A. (October 2006). The discovery of endogenous retroviruses. Retrovirology. [Online]. Available: www.retrovirology.com/content/3/1/67

[11] M. D. Lairmore, *et al*., Molecular Determinants of Human T-lymphotropic Virus Type 1 Transmission and Spread, Viruses, 2011.

[12] M. Matsuoka and K. T. Jeang, "Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation," *Nat Rev Cancer*, vol. 7, no. 4, pp. 270-280, Apr. 2007.

[13] R. Mahieux and A. Gessain, "The human HTLV-3 and HTLV-4 retroviruses: New members of the HTLV family," *Pathologie Biologie*., vol. 57, no. 2, pp. 161–166, 2009.

[14] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th International Conference on Very Large Data Bases*, Santiago, Chile, September 1994, pp. 487-499.

Yoondeok Jeon was born in 1996. He is currently a student of Hankook Academy of Foreign Studies, Republic of Korea with science major. He is deeply interested in biology and medical science, and likes to study about the structure of viruses. He especially have greatest interest in retroviruses, and have been published several journal articles about them within a year.

Jiwoo Oh is currently a student of Hankuk Academy of Foreign Studies, Republic of Korea. She is specialized in natural science programs with her strong interest to bioinformatics and medical science. She is actively studying virology and a remedy for retroviral infections for which there is still no validated cure. She wrote more than 5 papers concerning cures for several viral infections in a year and her enthusiastic work is continuing.

Yongha Jo is currently a student of Hankuk Academy of Foreign Studies, Republic of Korea with science major. She is deeply interested in biology and medical science, and most of her work is done by analyzing the structure of proteins using various algorithms including the apriori algorithm.

Taeseon Yoon is currently a teacher of Hankuk Academy of Foreign Studies. He teaches, and is a specialist in pattern analysis using multiple programs including Support Vector Machine, Neural Network, Decision Tree, etc.