Comparison of HIV-1, SIV-1, and SHIV in Protein Translation Using Apriori Algorithm

Junhyung Bae, Jongjun Lee, Yeji Jang, Soyoon Boo, and Sookyoung Lee
Hankuk Academy of Foreign Studies, Yongin, Gyeonggi, Republic of Korea
Email: hjweebee1@naver.com, ljj960819@gmail.com, yeji6032@naver.com, syboo4022@gmail.com, tazo1107@naver.com

***Abstract*—Retroviruses are viruses which have special ability to retro transcript their genetic data. HIV (Human Immunodeficiency Virus) which has been the cause for AIDS (Acquired Immunodeficiency Syndrome) in human is a kind of retroviruses. SIV (Simian Immunodeficiency Virus) and artificial SHIV (Simian-Human Immunodeficiency Virus) which are similar to HIV are lentiviruses. Since SIV do not cause any syndrome in simians, we expected that applying the genetic data of SIV and SHIV would assist finding new genetic cure for HIV-related diseases such as AIDS. Comparing the genetic sequences of these genes, we found the association among the viruses' genes. We used apriori algorithm for data mining process, and compared four genes (*env, nef, tat, vif*) which exists in all of three viruses' genomes. We focused on these four genes, for they have significant roles in viruses' recognition of host cell.**

***Index Terms*—bioinformatics, retrovirus, lentivirus, HIV**

I. Introduction

A. Viruses

Viruses cause diseases which are usually incurable with conventional antibiotics. One of the characteristic of viruses, not having cellular structure, makes them hard to cure, for most treatments for infectious diseases are specific to ones' cellular characteristics. They consist of either a RNA or DNA genome surrounded by a protective, virus-coded protein coat, the envelope. A complete virus particle is called a virion. The main function of the particle is delivering its DNA or RNA genome into the host cell so that the host cell can express their genome [1].

1) Retrovirus and lentivirus

Retroviruses are characterized by the two defining trademarks of replication: the reverse transcription of the genomic RNA into a linear double-stranded DNA copy, and the following covalent integration of this DNA into the host genome [2]. It also has a characteristic of infecting non-dividing cell such as T-lymphocytes. This trait helps retrovirus to become a good gene vector. Lentivirus is a genus of viruses of the Retroviridae family. Lentiviruses have a long life cycle the fact that means they can come out from the host cell long after the infection.

Manuscript received March 1st, 2014; revised May 12th, 2014.

2) Genetic analysis of retro and lentiviruses

There are 4 genomes that are necessary for retroviruses. They are *gag*, *pro*, *pol* and *env*. [3]. The *gag* gene encodes Gag protein, which is pivotal protein in viral assembly. Gag protein differs in each virus. Therefore, it functions as antigen determiner. The *pro* gene codes the viral protease. It is essential in facilitating the maturation of viral particles. The *pol* gene encrypts reverse transcriptase (RT), integrase, and other proteins that aids in proliferation of viruses. The gene *env* is in charge for the surface glycoproteins and transmembrane proteins which facilitate membrane fusion and cellular receptor binding.

There are 4 common genes that we used in the experiment, which exist in all of HIV-1, SIV-I, and SHIV. They are *env, nef, tat, vif.* The *nef* gene regulates Nef (Negative factor) protein. This protein reduces the level of CD4 of the cell's surface. Also, it induces non-dividing infected cell to divide. The *tat* gene regulates *Tat* protein, which activates transcription of the virus. The *vif* gene regulates Vif protein, which is believed to produce other virus particles [2].

3) Human Immunodeficiency Virus

HIV (human immunodeficiency virus) is a lentivirus that causes a continuous human immunodeficiency syndrome, which is also called AIDS (acquired immunodeficiency syndrome). When it infects the human, it exists in bodily fluids, such as semen, blood, or breast milk. It destroys human immune system by infecting helper T cells (specifically $CD4^+$ T cells), macrophages, and dendritic cells, which are essential in human immune system to prevent outer invasion. When infected, the host's $CD4^+$ T cell level declines to the critical level to prevent its own body from other infections, even though they are subtle [4].

It contains two single-stranded RNAs. These two copies of RNAs have genetic information on nine genes that the virus has. They are *gag*, *pol*, *env*, *tat*, *rev*, *nef*, *vif*, *vpr*, and *vpu* [5]. The most common virus among HIVs is HIV-1, which is believed to be derived from chimpanzees.

4) Simian Immunodeficiency Virus

Simian immunodeficiency viruses (SIVs) are retroviruses in at least 45 species of African primates. SIVs are believed to have longer history than HIVs, for HIVs are evolved from the SIVs. It is believed that SIV in simians had transmitted to humans through blood contact while hunting, resulting in SIVs' mutation to HIVs [6]. However,

doi: 10.12720/jomb.4.2.130-134

in contrast with HIV infections in humans, SIV infections in simians appear to be non-pathogenic in many cases, except some species that catch Simian AIDS when exposed to the virus [7].

5) Simian-Human Immunodeficiency Virus

Simian-Human Immunodeficiency Virus (SHIV) is a recombinant virus developed by humans. The similarity in the genetic composition and organization of HIV-1 and SIV makes it possible to construct recombinant viruses that show properties of both families. It can make a condition that looks like infection of simian with HIV-1. These viruses have been engineered to contain HIV-1 *env*, and they also contain genes which code reverse transcriptase.

The development of SHIV viruses has been especially advantageous because it provides the more relevant tool for research in studying properties of HIV-1 infection in a Non-Human Primitive setting. These properties include HIV-1 envelope characteristics that affect transmission and pathogenesis [8].

B. Experiment

As we earned noticeable benefits from developing SHIV, which contains both properties of HIV and SIV and compared and contrasted HIV and SIV's characteristics, we could identify the differences and similarities among HIV, SIV, and SHIV. Especially, when the viruses are analyzed genetically, we could make the procession from the very base of viruses' biological structure. Using predictive Apriori, we analyzed from nucleotides to amino acid protein sequence coded on four genes mentioned above and compared the genetic structure and the protein sequence of the viruses. As a result, we found similarities and some subtle differences in transcription and translation process.

C. Apriori Algorithm

Apriori is an algorithm that is frequently used in data mining process. Using transactional databases, the apriori algorithm find association rules in the given data. This algorithm is useful in finding how much data have exceeded threshold which user had set. Continuously making new seed item sets with pre-existing items in given threshold, association rules of data is compared. Among various kinds of data, DNA sequences are one of the appropriate datasets that apriori should be applied since DNA sequence data do not have timestamps.

II. Method

A. Dataset

Genetic information of viruses is from NCBI's database. To compare genetic makeup and nucleotide composition of HIV, SIV, and SHIV virus, complete genome was used. For HIV, NCBI Reference Sequence: NC_001802.1 was used for the analysis. Likewise, NCBI Reference Sequence: NC_001549.1 was used for the analysis of SIV virus, and NCBI Reference Sequence: NC_001870.1 was used for the analysis of SHIV virus. In these datasets, *env*, *nef*, *vif*, and *tat* genes have already been marked, so these genes could be used in the comparison between viruses.

B. Program

Predictive Apriori, which analyzed voluminous gene data via apriori algorithm, was used for the experiments. Predictive Apriori uses association rule algorithm. In this algorithm, two critical concepts, which are "support" and "confidence" is combined into a single concept, predictive "Accuracy". Predictive Apriori association rule algorithm uses "Accuracy" to generate the rules of Apriori association. Fig. 1 summarizes the Apriori algorithm explained above.

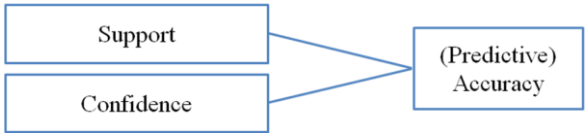


Figure 1. Apriori algorithm

This "accuracy", which is computed by using "support" and "confidence" data, appears in the results of the analysis of virus genes. Also, in Predictive apriori algorithm, when "n" is stated by the user of the program, the program draw a conclusion based on "n" best association rule.

III. Results and Discussion

A. Results

By using DNA sequence dataset and extracting meaningful areas, or exon, which encode amino acids and actually contribute in protein translation, we found significant similarities among three viruses. For each analysis, the results appeared as following:

1. amino1=R 8 ==> cleavage=yes 8 acc:(0.90768)
2. amino1=G 7 ==> cleavage=yes 7 acc:(0.89627)
3. amino5=L 7 ==> cleavage=yes 7 acc:(0.89627)

The results above are the analysis of vif protein of SHIV virus, using 7-window. "Amino1=R" means in the 7-window, the "R" amino acid takes the first space, and will be translated first among 4 other amino acids in the same window. The number "8" means there are 8 amino1=R that has transcended the limit that was set by the program. Accuracy is the feature that implies both "support" and confidence. In the results of the data, amino acids, locations, and the numbers are arrayed according to the "accuracy"

Analyzing the results of the DNA sequence dataset analysis, top one-hundred amino acids-location data was used. *env* showed accuracy range 0.85 to 1.0 and *nef* showed accuracy range 0.6 to 1.0. Also, *tat* showed accuracy range 0.80 to 1.0 and *vif* showed accuracy range 0.55 to 1.0. Higher accuracy means the results are more reliable.

Using the results, amino acids and its following numbers were counted. Then, the arranged data was shown in a graph. Vertical lines between the highest spot and the

lowest spot were drawn for each encoded amino acids, in order to identify differences among three viruses.

Among the proteins several characteristics were found and each protein showed different characteristic. Following graphs from Fig. 2 to Fig. 13 are the results.

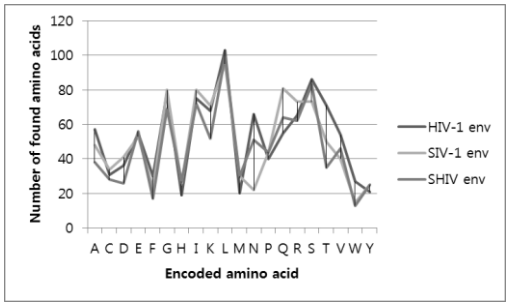


Figure 2. Env-protein 5-window predictive apriori result

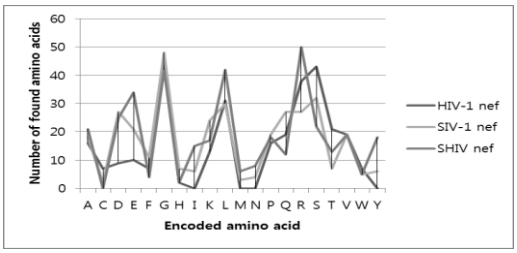


Figure 3. Nef-protein 5-window predictive apriori results

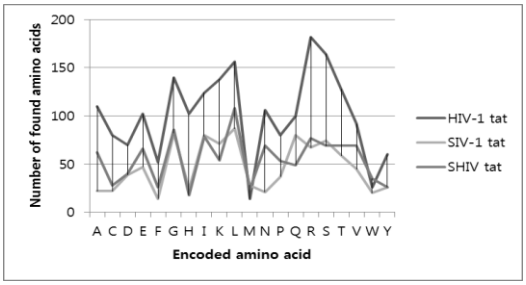


Figure 4. Tat-protein 5-window predictive apriori results

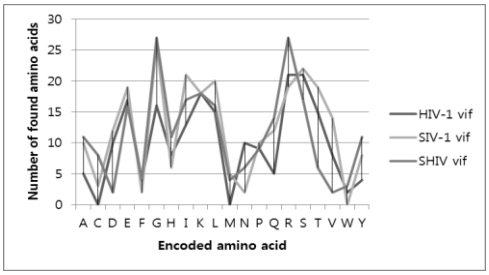


Figure 5. Vif-protein 5-window predictive apriori results

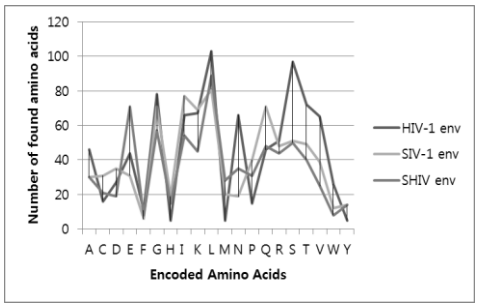


Figure 6. Env-protein 7-window predictive apriori results

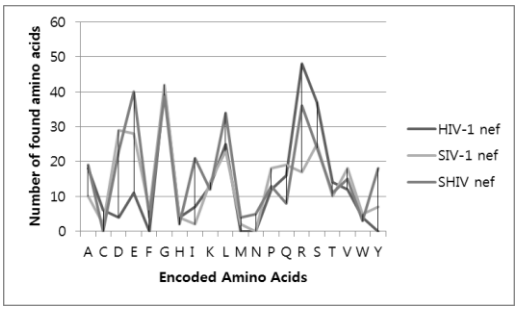


Figure 7. Nef-protein 7-window predictive apriori results

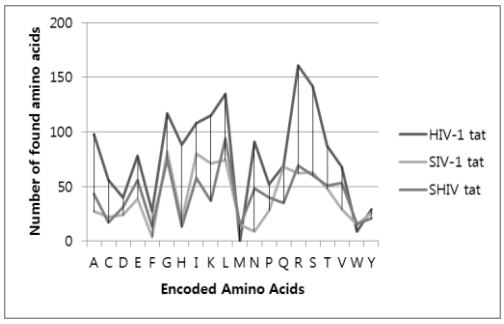


Figure 8. Tat-protein 7-window predictive apriori results

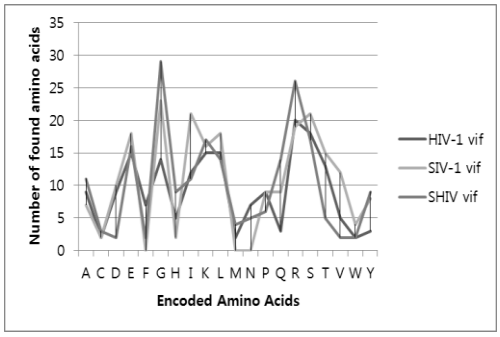


Figure 9. Vif-protein 7-window predictive apriori result

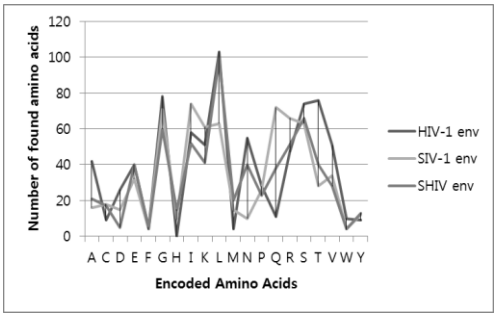


Figure 10. Env-protein 9-window predictive apriori results

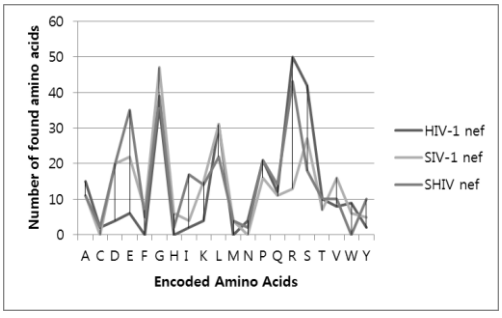


Figure 11. Nef-protein 9-window predictive apriori results

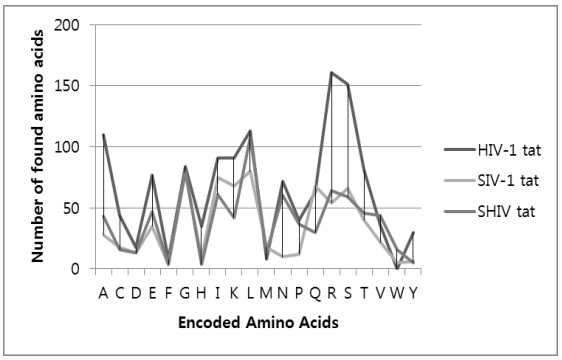


Figure 12. Tat-protein 9-window predictive apriori results

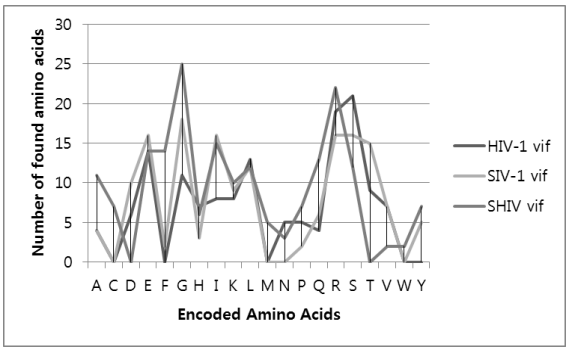


Figure 13. vif-protein 9-window predictive apriori results

B. Discussion

HIV, SIV, and SHIV virus showed noticeable similarities and differences on 5-Window analysis

As we can see in the Fig. 2, among four different proteins, *env* protein showed the greatest similarities, though three viruses showed differences on some encoded amino acids such as T, N, and Q. *env* protein encodes surface glycoprotein, and transmembrane proteins. Also it determines the capacity of cell viruses to penetrate cell membrane and infect, since glycoprotein on the surface of the viruses fuse with the acceptors on cells' membrane and enables viruses to permeate the cell's membrane. Similarities in the amino acids that are decoded to form env proteins mean three viruses have similar glycoprotein, which will explain why HIV and SIV viruses both show strikingly similar genetic makeup particularly on primates.

Also, taking *env*-protein in consideration, SHIV virus show approximately average number for each encoded amino acids. This result can be explained by the fact that SHIV was made manually, through the synthesis of HIV gene and SHIV gene artificially. This result also explains the reason that SHIV viruses show particularity on simians by presenting noticeable lethality and contagiosity.

As we can see in the Fig. 3, while *nef* protein showed some similarities, noticeable differences were found on D, E, R, S, and T encoded amino acids. These genomic differences might contribute to particularities of three viruses. The differences in negative factor protein between three viruses are induced by this difference. Like *env* protein, SHIV showed average number between HIV and SIV virus for each amino acids, which can be the evidence that SHIV is the result of genomic fusion between HIV and SIV viruses. *vif* protein also showed differences on A, C, G, T, V encoded amino acids.

As we can see in the Fig. 4, in *tat* protein, every amino acid except M and W showed greatest differences. The gap between the highest and the lowest point was the largest among four proteins. However, the tendency of the graph between three viruses was noticeably similar. Also, in most amino acids the HIV showed the highest number, and SIV showed the lowest number, and SHIV was more similar to SIV, judged by the graph. This result will imply that SHIV is more similar to SIV in the makeup of *tat* protein.

As we can see in the Fig. 5, *vif* protein showed almost equal percentage of similarities that was shown in *tat* protein, but they were not as similar as *env* protein.

7-Window and 9-Window analysis showed similar tendency

As we can see in Fig. 6 to Fig. 9 above, on 7-window analysis, HIV, SIV, and SHIV virus showed similar tendencies with 5-Window analysis. The analyzed results of *env*, *nef*, and *vif* proteins were similar in every virus. However, compared to 5-window analysis, both 7-window analysis and 9-window analysis showed certain differences on certain encoded amino acids.

On 7-window analysis of *env* protein, T, N, and Q encoded amino acids showed one of the greatest differences among other amino acids. Moreover, new amino acids, such as V and W, showed differences, as it is shown by vertical lines, which was not observable in 5-window analysis. Oher proteins such as *nef*, *tat*, and *vif* showed similar tendencies with *env* protein. On every protein, the gap between the highest point and the lowest became wider than 5-window analysis. Also, some amino acids were newly found as showing differences on each virus. However, among every protein, *tat* protein still have shown the largest gap, and SHIV virus still had approximately average values on every amino acids, as it was shown on 5-window analysis.

Also, as we can see in Fig. 10 to Fig. 13 above, on 9-window analysis, the results of apriori analysis have shown similar tendency with those of 7-window analysis, except *tat* protein. As shown in the Fig. 13, the graph of *tat* protein, the gap between viruses became narrower than that of 5-window and 7-window analysis. Still, except *tat* protein, *env*, *nef*, and *vif* protein have shown similar tendency, with wide gap between the highest and lowest point on each protein even though there were a few newly found amino acids.

There are certain tendencies that as the number "n" in "n"-window analysis is ascended the graphs tend to have wider gaps. This phenomenon can be explained by the characteristic of Apriori Analysis we used. As the number "n" in "n"-window increases, the criterion of the apriori analysis become stricter.

Comparison to other references

These results, which showed recognizable similarities between HIV, SIV, and SHIV viruses, match with other precedent studies in some extent. It has been validated that SIV virus is related to HIV in the aspect of the antigenicity of proteins and its properties. Also, it was identified that

supposedly, due to its similarities to HIV virus, SIV induces a disease that has remarkable similarity to human AIDS in the common rhesus macaques [9].

Our results which show that HIV, SIV, and SHIV viruses have similarities on envelope proteins can also be validated by a precedent research on sensibility of these viruses to various types of drugs. SHIV virus, which is recombinant of HIV and SIV, is susceptible to enfuvirtide, for it expresses HIV-1 envelope glycoproteins. Also, the SHIV virus which contains an HIV-1 RT, NNRTI will be used as a coping material to SHIV virus infections [10].

IV. Conclusion

In the process of researching and developing vaccines for HIV which causes serious diseases for human, scientists have been frustrated, for there has not been any effective means to cure the disease because of its complexity in its life cycle and its long incubation period. After analysis using apriori algorithm with three other types of windows, we thoroughly investigated genetic structure of 4 common genes among HIV(HIV-1), SIV(SIV-1), and SHIV. As a result of the analysis, we found that information contained in three viruses seams similar but has significantly different sequences. Since SIV does not cause vital disease to simians as HIV does to humans, we look forward to finding new ways of curing HIV using our results. We strongly believe that these results will be helpful on medical society.

Acknowledgment

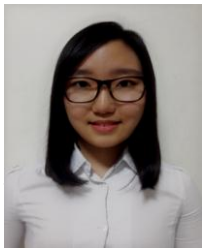
Special thanks to Taeseon Yoon, who lead all of us to the world of bioinformatics.

References

[1] S. Baron, *Medical Microbiology*, 4th Ed. University of Texas Medical Branch at Galveston, 1996.
[2] J. M. Coffin, S. H. Hughes, and H. E. Varmus, *Retroviruses*, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 1997.
[3] J. Leis, D. Baltimore, J. M. Bishop, J. M. Coffin, E. Fleissner, S. P. Goff, S. Oroszlan, H. Robinson, A. M. Skalka, H. M. Temin, and P. K. Vogt, "Standardized and simplified nomenclature for proteins common to all retroviruses," *J. Virol*, vol. 62, pp. 1808–1809, 1988.
[4] A. Cunningham, H. Donaghy, A. Harman, M. Kim, and S. Turville, "Manipulation of dendritic cell function by viruses," *Current Opinion in Microbiology*, vol. 13, no. 4, pp. 524–529, 2010.
[5] C. Kuiken, T. Leitner, B. Foley, B. Hahn, *et al.* "HIV sequence compendium 2008," *Theoretical Biology and Biophysics*, 2008.
[6] M. Peeters, V. Courgnaud, and B. Abela, "*Genetic diversity of lentiviruses in non-human primates*," *AIDS Reviews*, vol. 3, pp. 3–10, 2001.
[7] H. Kestler, T. Kodama, D. Ringler, *et al.* "Induction of AIDS in rhesus monkeys by molecularly cloned simian immunodeficiency virus," *Science*, vol. 248, no. 4959, pp. 1109–1112, 1990.
[8] L. E. Pereira, P. Srinivasan, and J. M. Smith, "Simian-human immunodeficiency viruses and their impact on non-human primate models for AIDS," *Immunodeficiency*, ch. 15, 2012.
[9] L. Chakrabarti, *et al.* "Sequence of simian immunodeficiency virus from macaque and its relationship to other human and simian retroviruses," *Nature*, vol. 328, no. 6130, pp. 543-547, 1987.
[10] M. Witvrouw, *et al.* "Susceptibility of HIV-2, SIV and SHIV to various anti-HIV-1 compounds: Implications for treatment and postexposure prophylaxis," *Antiviral Therapy 9.1*, pp. 57-66, 2004.

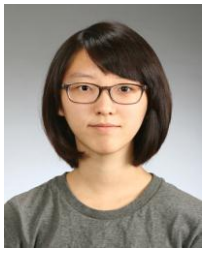
Junhyung Bae was born in 1996. He is currently a student in science major of hankuk academy of foreign studies. He is interested in bio-science and applying bio-informatics to analyzing gene and finding new facts about human genes.

Jongjun Lee was born in 1996. He is currently a student in science major of hankuk academy of foreign studies. He is interested in bio-science and applying bio-informatics to analyzing gene and finding new facts about human genes.



Yeji Jang was born in 1997. She is currently a student in international studies major of hankuk academy of foreign studies. To extend her passion for biology, she applied bio-informatics to her research, analyzed genomic sequences and found new facts about genes.

Soyoon Boo was born in 1996. She is currently a student in science major of hankuk academy of foreign studies. She is interested in bio-science and medical science and applying bio-informatics to analyzing gene and finding new facts about human genes.



Sookyoung Lee was born in 1996. She is currently a student in science major of hankuk academy of foreign studies. She is interested in bio-science and medical science and applying bio-informatics to analyzing gene and finding new facts about human genes.