Analysis of Structural Relationship between Immunodeficiency Viruses Using Support Vector Machine

Yoondeok Jeon, Jiwoo Oh, Seungjae Lim, Yewon Choi, Sungmoon Kim, and Taeseon Yoon

Abstract—**Immunodeficiency viruses incur the lack of ability to fight against pathogens and weaken one's immune system. They all belong to a genus lentivirus. Since these viruses do reverse transcription that lacks normal proofreading of DNA transcription, they mutates very often. The materials used in the research include several viruses which could be mutated from common ancestor; they are SIV, HIV, BIV, and FIV. Focusing on the four viruses, we conducted two main experiments using their RNA sequences by munticlass SVM; for one, analysis of genetic properties by identifying each one's structures. For two, comparison of the structural relationship between a couple of viruses after setting six subsets from four. In the first one, the structure of all the four immunodeficiency viruses presented nonlinearity. In the second one, the structural relationship between SIV and FIV was nonlinear (RBF function) with extraordinarily high accuracy. While all viruses showed nonlinear structures, comparison of SIV and FIV verified us their genetic relationships of nonlinearity. The genetic similarity of the four viruses supports a notion that they are rooted from a common ancestor.**

Index Terms—**Immunodeficiency virus, mutation, structural relationship, support vector machine (SVM), nonlinear, common ancestor.**

I. Introduction

Immunodeficiency is literally, a state in which the ability of an immune system is compromised or completely absent. Most cases of immunodeficiency is acquired (opposite to innate) by immunosuppressive agents, pathogenic bacteria or viruses. Also, certain diseases directly or indirectly cause immunocompromised state that a person has got his immune system weakened [1]. These diseases include Acquired Immunodeficiency Syndrome (AIDS) and cancer (especially developing from bone marrow or blood cells) [2].

Particularly, diseases caused by retroviral infections are generally hard to cure or entirely incurable. The incurability is mainly derived from the virus's own feature of highly frequent mutations caused by reverse transcription, a completely reverse process to usual DNA transcription. Retroviruses use reverse transcriptase (RT) to convert RNA to DNA for replications [3]. Since this procedure is extremely unstable and error-prone, and these properties lead to incurability and drug resistance [4].

'Lentivirus' is a genus of viruses and belongs to family Retroviradae. The prefix "Lenti" comes from the Latin for "slow." Slow incubation period makes this genus of viruses special from other retroviruses [5]. The species include HIV, SIV, and visna.

This research focuses especially on immunodeficiency viruses, all belonging to genus lentivirus. Human Immunodeficiency Virus (HIV), known to cause AIDS, is believed to have origin

nated through the evolution of Simian Immunodeficiency Virus (SIV). Including these, we chose four immunodeficiency viruses, HIV, SIV, BIV (bovine-), and FIV (feline-). We expected that these viruses would show common characteristics in its genetic structures and properties. Hence, the research is mainly based on this hypothesis, done by analyzing their DNA sequences using Support Vector Machines.

II. Materials and Methods

A. Virus

HIV-1(Human immunodeficiency virus type 1) is the most well-known among all lent viruses and as mentioned above, brings AIDS to human being [6], [7]. Victims of AIDS experience progressive failure of an immune system, especially the loss of function of helper T cells, and the body grows vastly vulnerable to other inflectional diseases. HIV is roughly spherical in structure with a diameter of 120 nm, relatively big from other viruses [8], [9]. It is composed of 2 copies of positive single-stranded RNA, and each single-stranded RNA is covered with nucleocapsid proteins, p7, and enzymes needed for the development of the virion. Unlike others, HIV shows fast replication cycle, with a generation of about 10^{10} virions every day, and this results in very high genetic variability [10], [11].

Simian immunodeficiency viruses (**SIVs**) can infect at least 45 species of non-human primates [12], [13], and it is notable that SIV sooty mangabeys (SIVsmm) and SIVchumpanzees (SIVcps) are to have crossed the species barrier and evolved to HIV. The virus is distinguished from other viruses in some part [14]. SIV and related retroviruses have variants of the protein TRIM5α in humans and non –human primates. APOBE3G/3F is also an important protein in restricting cross-species transmission.

Bovine immunodeficiency virus (BIV), with size of 110-130nm in average, fatally damages the immune system of a cattle [15]. It can be easily transmissed by the exchange of bodily fluids. Lymphocytes, monocytes and macrophages are mainly infected by the virus. Accordingly, leukocytosis and lymphadenopathy are its early infection symptoms [16]. Most symptoms of BIV infection are similar to those of AIDS. Since it can affect non-dividing cells, the virus can be used in gene therapy. Also, as a non-primate virus, it does not

Manuscript received January 24, 2014; revised March 24, 2014.

The authors are with the Hankook Academy of Foreign Studies, Republic of Korea (e-mail:{junyd5469, ojw0414, tonylim0930, choiyewon961001, p0q1013}@gmail.com, tsyoon@hafs.hs.kr).

damage human cells and property enables BIV a safer method for gene therapy [17].

Feline immunodeficency virus (FIV) affects the immune system of feline species including a cat and causes AIDS-like syndrome [18], [19]. Similar to HIV [20], FIV infects many immune cells including CD4+ and CD8+ T lymphocytes, B lymphocytes, and macrophages, crucially leading CD4+ T lymphocytes(or T-helper cells) to debilitate and diminish, which play an important role in activating other immunocytes. As the virus progresses, its enveloping glycoproteins interacts with the receptors of the target cells, binding to the receptor CD134 on the host cell. This interaction brings about the fusion of viral and cell membrane, and the viral RNA transfers into the cytoplasm. This is followed by reverse transcription and integration into the cell genome, resulting in the infection of FIV [21].

B. *SVM* (*Support Vector Machine*)

Support vector machines (SVM) are supervised learning models including algorithms that analyze data, recognize patterns, and these results are used for classification [22]. Basically, SVM implements prediction of data by learning how to classify two types of data. When different data sets are given, the SVM finds a hyperplane that efficiently classifies the data sets. By analyzing the data set, several hyperplanes that divide the data can be found. To find the most accurate hyperplane, the SVM finds the one that produces the largest margin between different data sets, since larger margin leads to low generalization error of the classifier. For example, in Fig. 1 and 3 hyperplanes are presented. H_1 does not effectively classify the data, and H_2 produces smaller margin compared to H_3. H_3 accurately divides the data into two, and it produces the largest margin between them, so it will be considered as the maximum-margin hyperplane, which will be the most reasonable classifier for these data sets. The two opposite-side-located samples which are the closest to the maximum-margin hyperplane and thereby used for calculating margin are called support vector. In Fig. 1, two samples in square boxes each are support vectors of SVM.

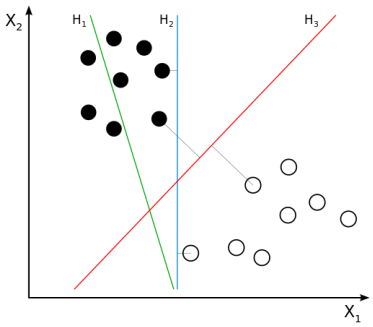


Fig. 1. Classification method of SVM.

With a given sets of data, the SVM algorithm randomly divides the data into two categories: training data and test data, and the training examples are given in order to train the model. To observe the results in better accuracy, the training of the models needs to be repeated several times while the components of each data set (training data set and test data set) are changed. This process is called cross-validation [23].

At first, linear SVM was used to classify the data, but it was soon realized that few data could be separated with linear method [24], [25]. Accordingly, accurate non-linear classification method was required, and the kernel method was adapted to SVM to solve the problem. Using kernel functions, inputs are mapped into higher dimensional spaces so that data can be linearly separated into two different spaces by new hyperplanes, which may be nonlinear in the original input space. Fig. 2 shows the simple example of how these kernels function, separating nonlinear data in a linear way. Furthermore, Fig. 3 explains the way how using radical basis function kernel (RBF kernel) enables data to be separated linearly in 3-dimension space, which was not able to be linearly classified in 2-dimension space. After the utilization of mapping method using various kernel functions in SVM, SVM could be an effective algorithm to perform non-linear classification.

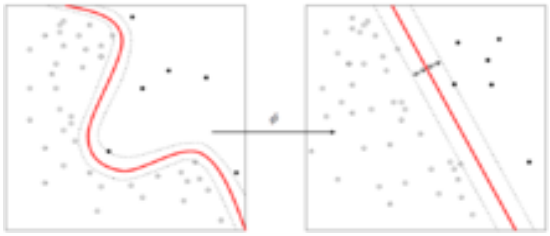


Fig. 2. Kernel machine.

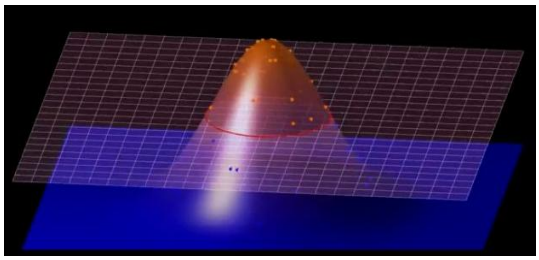


Fig. 3. Non-linear classification by using RBF Kernel.

Including RBF kernel, the most commonly used kernels in SVM are normal sigmoid and polynomial kernel. To simply introduce them, below are the definitions of each kernel;

Normal kernel:

$$k(x,y) = x^T y + c \tag{1}$$

Sigmoid kernel (also known as hyperbolic tangent kernel):

$$k(x,y) = \tanh\left(ax^T y + c\right) \tag{2}$$

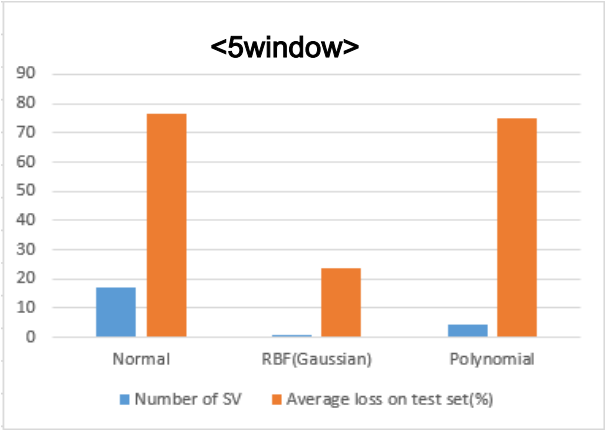
Radical basis function kernel (also known as RBF kernel, Gaussian kernel):

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{3}$$

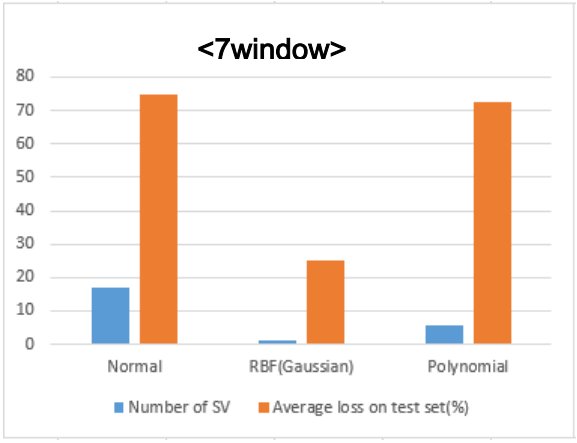
Polynomial kernel:

$$k(x,y) = \left(ax^T y + c\right)^d \tag{4}$$

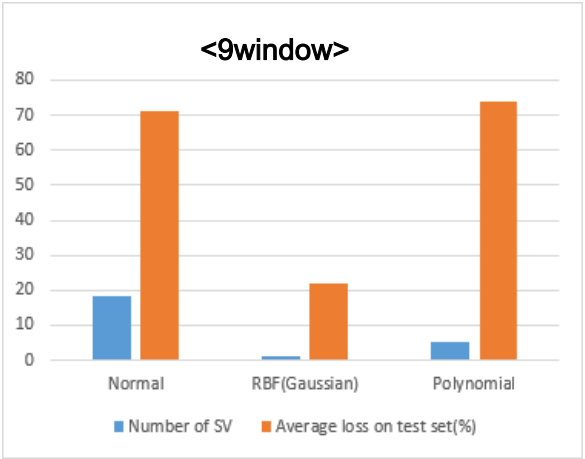
III. Experiment



	Normal	RBF	Polynomial
Number of SV	17	1	4.3
Average Loss	76.76137	23.63636	74.88636



	Normal	RBF	Polynomial
Number of SV	16.8	1	5.6
Average Loss	75	25	72.3333



	Normal	RBF	Polynomial
Number of SV	18.5	1	5.3
Average Loss	71.087	22.0652	74.0218

Fig. 4. Classification result using multiclass SVM.

In this paper, multi-class support vector machine ($SVM^{multiclass}$) is used, which uses the multi-class formulation [26]. For all experiments, 10-fold cross-validation was adopted for higher accuracy, and three kernel functions was used for classification; Normal, RBF, and Polynomial. The data was divided into 10 different sets. One data set was used as a test data, and others were considered as training data sets, and the models were trained with these. Thus, 10 data sets were all used as test data for one time respectively, thus 10 experiments were made.

The Fig. 4 is the result of classification using $SVM^{multiclass}$.

For each window, 10 experiments were done. The number of support vectors used to classify the data and the percentage of average loss on test set were all different in each experiment. Thus, we looked for the average value of them and made charts. Fig. 4 is the result of the analysis, and y-axis represents the average values in percentage terms.

The result shows that the number of support vectors used by $SVM^{multiclass}$ is relatively low when RBF kernel was implemented, compared to results derived by the implementation of normal and polynomial kernel. Furthermore, the average loss on test sets was also low with RBF kernel. The loss refers to the discordance with the kernel. Therefore, high percentage of loss with certain virus does not accommodate the overall structure of it. In Fig. 4 overall, RBF kernel produces the smallest loss percentage while other kernels produce high percentage of average loss. By applying the characteristics of each kernel to this result values, it can be inferred that the overall structure of the viruses present nonlinearity.

Since multiclass SVM collects all the data given and produces the result using them, there is a possibility that all the data is mixed, making the process become error-prone. Therefore, to identify the structural relationship in a deeper level, 6 different datasets comparing two different viruses were made; they are (HIV-1, SIV), (HIV-1, BIV), (HIV-1, FIV), (SIV, BIV), (SIV, FIV), and (BIV, FIV). SVM was used to find the structural relationship between two viruses. The results are in Fig. 5 and Fig. 6 below.

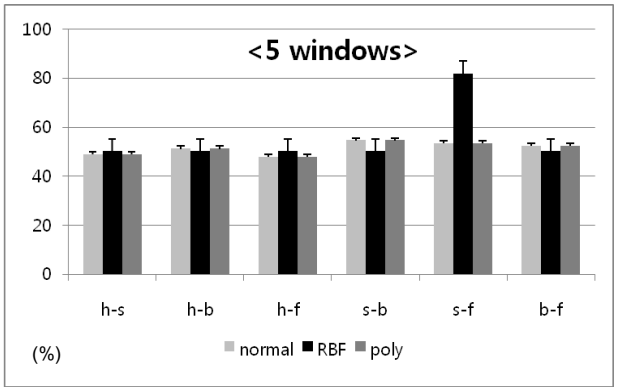


Fig. 5. Classification result of 6 datasets of 5 window.

This chart indicates the relationship between two viruses (total 6 couples). Their RNA sequences were analyzed in 5 window with 3 different kernel functions.

Fig. 5 shows the result of analysis of RNA sequences in 5 window. The y-axis in the chart (Fig. 5) indicates classification accuracy between two viruses, using three

different kernel functions. The results were analyzed in percentage terms. Other than any two couples, a relationship between SIV and FIV classified with RBF function shows extraordinarily high accuracy. This validates the unique characteristics of RBF function, presenting that RNA sequences of SIV and FIV are quite similar to the shape of Gaussian function, therefore "non-linear." Moreover, the lower the number of window is, the higher the classification accuracy goes. Compared to the analysis in the data divided into 7 and 9 window, 5 window would perform with advanced credibility.

In the analysis in 7 window, there was no conspicuous difference from that in 5 window, showing that classification accuracy between SIV and FIV was far above the average (data not shown).

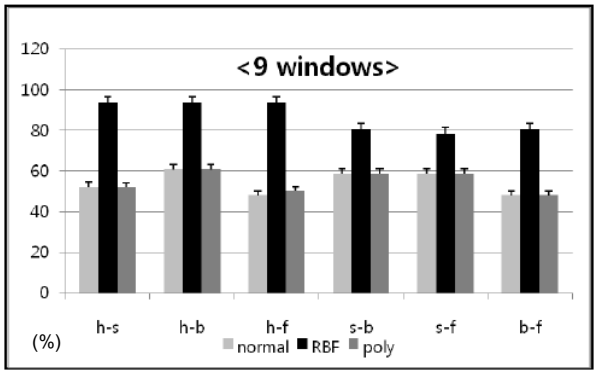


Fig. 6. Classification result of 6 datasets of 9 window.

This chart indicates the relationship between 6 couples of two viruses. Their RNA sequences were analyzed with 3 different kernel functions.

The outcome from 9 window revealed a dissimilar tendency. The classification accuracy in all couples was measured high in only RBF function. This result is in fact, quite opposite to the previous. According to the fact that classification accuracy of SIV/FIV was significantly higher than any other couples, the accuracy value of SIV/FIV here analyzed in 9 window shows relatively lower accuracy than other couples of RBF function (the black one). This disparity is due to the difference in classification performance of each window. As mentioned earlier, higher number of window represents inferior quality of classification accuracy. Hence, the close relationship between SIV and FIV disproves the result presented in Fig. 6.

IV. Conclusion and Discussion

We used multiclass support vector machine to understandthe overall genetic structural tendency among 4 retroviruses, HIV, SIV, BIV, and FIV. The result of the experimentrevealed the lowest percentage of average loss on the test setwhen RBF (Gaussian) kernel was used, and the number ofsupport vectors was also the lowest. If the average losspercentage is high, it means that the data cannot be separatedaccurately. Since the average loss percentage of normal orpolynomial kernel with exponent 1 was high enough, it can be inferred that the overall structure for immunodeficiencyviruses is not linear. Conversely, as the loss value of RBFkernel was relatively low, it is obvious that the virus showsstrong non-linearity.

We've done further experiments to gain deeperunderstanding of the structural relationship between theviruses. We divided the viruses into 5, 7, 9 windows and made6 sets of data containing 2 different viruses, and appliednormal, RBF, and polynomial kernel to compare thestructures. The result showed low accuracy, lower than 70percent, for classification by most applied kernels. When theamino acid sequences of each virus was divided into 5window, only SIV and FIV was accurately divided when RBFkernel was used, and for 7 window, HIV-1 and SIV withnormal kernel and polynomial kernel. Exceptionally, whenthe sequences were divided into 9 window, in all cases, theaccuracy for classification was above 80% when RBF kernelwas applied. Yet, in the experiment with 9 window, there lay ahigh possibility of error since the number of components ineach unit get bigger. Thus, SIV and FIV could show evidentdifference in structure even they both have strongnon-linearity.

As the SVM acts as a classification model, it cannot classifycorrectly if the different data sets are extremely similar.Considering this fact, the result of the experiment proves thesimilarity in structure of all immunodeficiency viruses. In thefield of classification, the accuracy of 50% is considered verylow, which means that the classification couldn't be done withease. In the experiment, most of the kernels showed around50% in accuracy. Even when polynomial was used withexponent 2 kernel to compare HIV-1 and SIV, the accuracy didn't show conspicuous difference (data not shown). Thisstrongly supports the assumption that HIV and SIV wouldshow similarity in structure. In conclusion, the experiment represents that all immunodeficiency viruses might havesimilar genetic structure. This notion especially supports thefact that SIVsmm in sooty mangabeys and SIVcpz inchimpanzees served as transmitter of the virus to humanspecies, resulting in mutation to HIV-1 and HIV-2. Furthermore, all immunodeficiency viruses might be derivedfrom common virus due to its structural and functionalsimilarities.

References

[1] S. Greenberg, *Immunodeficiency*, University of Toronto, February 5, 2009.
[2] A. K. Abbas and A. H. Lichtman, *Basic Immunology: Functions and Disorders of the Immune System*, 3rd Ed., Saunders/Elsevier, 2010.
[3] R. Kurth and N. Bannert, *Retroviruses: Molecular Biology, Genomics and Pathogenesis*, UK: Caister Academic Press, 2010.
[4] P. Medstrand, L. van de Lagemaat, C. Dunn, J. Landry, D. Svenback, and D. Mager, "Impact of transposable elements on the evolution of mammalian gene regulation," *Cytogenet Genome Res*., vol. 110, issue 1-4, pp. 342-352, 2005.
[5] J. R. Gilbert and W.-S. Flossie, "HIV-2 and SIV vector systems," in *Lentiviral Vector Systems for Gene Transfer*, G. L. Buchschacher, Ed. Georgetown, TX: Eurekah.com, 2003.
[6] R. A. Weiss, "How does HIV cause AIDS?" *Science*, vol. 260, no. 5112, pp. 1273-1279, May 1993.
[7] D. C. Douek, M. Roederer, and R. A. Koup, "Emerging concepts in the immunopathogenesis of AIDS," *Annu. Rev. Med.*, vol. 60, pp. 471-484, 2009.
[8] 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.
[9] B. Fisher, R. P. Harvey, and P. C. Champe, *Lippincott's Illustrated Reviews: Microbiology (Lippincott's Illustrated Reviews Series)*, Hagerstown, MD: Lippincott Williams & Wilkins, 2007.

[10] D. L. Robertson, B. H. Hahn, and P. M. Sharp, "Recombination in AIDS viruses," *J Mol Evol.*, vol. 40, no. 3, pp. 249-259, 1995.
[11] A. Rambaut, D. Posada, K. A. Crandall, and E. C. Holmes, "The causes and consequences of HIV evolution," *Nature Reviews Genetics*, vol. 5, pp. 52-61, January 2004.
[12] M. Peeters, V. Courgnaud, and B. Abela, "Genetic diversity of lentiviruses in non-human primates," *AIDS Reviews*, vol. 3, no. 1, pp. 3-10, 2001.
[13] M. Peeters and V. Courgnaud, "Overview of primate lentiviruses and their evolution in non-human primates in Africa," in *HIV Sequence Compendium.*, C. Kuiken, B. Foley, E. Freed *et al.*, Ed. Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2002, pp. 2-23.
[14] H. Kestler, T. Kodama, D. Ringler *et al.*, "Induction of AIDS in rhesus monkeys by molecularly cloned simian immunodeficiency virus," *Science*, vol. 248, pp. 1109-1112, 1990.
[15] M. Horzinek, L. Keldermans, T. Stuurman, J. Black, A. Herrewegh, P. Sillekens, and M. Koolen, "Bovine immunodeficiency virus: Immunochemical characterization and serological survey," *Journal of General Virology*, vol. 72, no. 12, pp. 2923-2928, 1991.
[16] M. C. St. Louis, M. Cojocariu, and D. Archambault, "The molecular biology of bovine immunodeficiency virus: A comparison with other lentiviruses," *Cambridge Journals Online*, pp. 125-143, 2004.
[17] R. Berkowitz, Y. L. Wei, K. Eckert *et al.*, "Construction and molecular analysis of gene transfer systems derived from bovine immunodeficiency virus," *J. Virol.*, vol. 75, no. 7, 2001.
[18] V. M. Lara, S. A. Taniwaki, and J. P. A. Júnior, "Occurrence of feline immunodeficiency virus infection in cats," *Ciência Rural*, vol. 38, no. 8, 2008.
[19] J. Richards, "Feline immunodeficiency virus vaccine: Implications for diagnostic testing and disease management," *Biologicals*, vol. 33, issue 4, pp. 215-217, 2005.
[20] J. H. Elder, Y. C. Lin, E. Fink, and C. K. Grant, "Feline immunodeficiency virus (FIV) as a model for study of lentivirus infections: parallels with HIV," *Curr HIV Res.*, vol. 8, issue 1, pp. 73-80, Jan. 2010.
[21] K. Hartmann, "Clinical aspects of feline immunodeficiency and feline leukemia virus infection," *Veterinary Immunology and Immunopathy*, vol. 143, issue 3-4, pp. 190-201. 2011.
[22] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, 1995.
[23] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification (technical report)," Department of Computer Science and Information Engineering, National Taiwan University, (2003).
[24] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821-837, 1964.
[25] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annual ACM Workshop on COLT*, Pittsburgh, PA: ACM Press, 1992, pp. 144-152.
[26] K.-B. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," in *Proc. the Sixth International Workshop on Multiple Classifier Systems*, 2005, vol. 3541, p. 278.

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.

Seung Jae Lim was born in 1996. He is currently a student in science major of Hankuk Academy of Foreign Studies, Korea. He is mostly interested in chemistry and biology. He has been studying pattern analysis and computer programming and its application to chemistry and biology.

Yewon Choi was born in 1996. She is currently a student of Hankuk Academy of Foreign Studies, Republic of Korea. She is deeply interest in studying virology and the structure of viruses. She is actively studying virology and applies the methods of bioinformatics in her studies.

Sungmoon Kim was born in 1996. She is currently a student of Hankuk Academy of Foreign Studies, Republic of Korea. She is deeply interested in studing viruses, and most of her work is done by analyzing the sequence patterns of them.

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.