Analysis of Banana Viruses with Decision Tree

Seohee Lee*, Wonil Roh, Taeseon Yoon
Department of Natural Science, Hankuk Academy of Foreign Studies, Yong-in, Republic of Korea.

* Corresponding author. Tel.: +82-10-8459-3566; email: orikkang0812@naver.com
Manuscript submitted January 5, 2015; accepted August 18, 2015.
doi: 10.17706/ijcce.2016.5.4.253-259

Abstract: As panama disease has appeared and damaged banana plantation, it posed serious threat to bananas, even leading some species to the brink of extinction. It is to find the solution to cure banana disease by disclosing different features of banana viruses. We analyzed 12 banana viruses : bract mosaic virus, mild mosaic virus, streak virus straubAcuminata Vietnam, CA streak virus, GF streak virus, IM streak virus, OL streak virus, VN streak virus, UM streak virus, UL streak virus, UI streak virus, UA streak virus. Sequences and amino acids of 12 viruses were analyzed and classified into groups by using Decision tree method. Through this research, we hope to find an efficient method to cure banana disease.

Key words: Banana virus, decision tree, banana streak virus, classification, bio-informatics.

1. Introduction

As panama disease, especially the TR4 virus, has appeared and run rampant at a rapid rate, the number of Cavendish banana is said to decrease drastically in the next few years. Cavendish banana hold almost 95% of international banana trade and this phenomenon is a critical threat to the agricultural market. Nowadays, not only TR4 virus but many different kinds of viruses, especially the Banana streak virus, threaten bananas. Thus, we have decided to investigate about the genetic characteristics of bananas differentiated by species and the viruses that cause diseases with bananas [1]. We will deal with these 12 Banana streak viruses and Banana mosaic viruses: Bract mosaic virus, Mild mosaic virus, Streak virus straubAcuminata Vietnam, CA streak virus, GF streak virus, IM streak virus, OL streak virus, VN streak virus, UM streak virus, UL streak virus, UI streak virus, UA streak virus [2].

We have found some other research that is related to Banana streak virus, and then make clear the differences and similarities between other studies and ours in terms of method and purpose of the study. Methodical way is different from other studies that most of them used PCR-based diagnostic method [3], [4], while bio-informatics and decision tree is used in this research [5]. Also, there are clear differences in a regard that we deal with precisely different types of banana streak viruses, such as Banana streak IM virus, Banana streak CA virus, Banana streak UM virus, etc. [6].

2. Materials and Methods

2.1. About Decision Tree

Decision tree is one of the methods which are commonly used in data mining. It detects targeted variables based on input variables. Decision tree is traditional solution to get the most efficient strategy leading to a goal. The manual form of the tree is drawn with leaves and branches. Different criteria applied on each point that 'branches' spread out. However, in this research, it is applied in the form of informatics [7], [8].

2.2. Method

Differences between windows are in the range of DNA which is cut to progress the experiment. Windows used for the experiment are three types: 17 window, 13 window, and 9window. Base sequences of 12 viruses were obtained from the NCBI [9], and because different windows means different standard how they cut sequences, experimental results also partially different between windows. 'Fold' is a sub-concept of window, and virus that used for standard of the experiment is different in each folds. Several folds are similar to the meaning of the several times of experiment.Computer programming works efficiently when it is applied in organizing Decision Tree by hastening the branching-out process of the graph. The experiment is preceded by comparing different class to a standard class in each fold. We can attain credible information quickly because the experiment was repeated many times and have high accuracy. The experimental results suggest rules and the frequency of these rules. Rule means notable tendency that the same particular amino acids appeared in the same position of two different classes, and frequency means accuracy of the rule. The specific rules we choose were filtered out from approximately 7000 rules obtained from experiment, and all of them represent the probability of at least 75%.

3. Classification of Classes

3.1. Classification

Table 1. Classes and Viruses' Name

Class	Virus	Class	Virus
Class1	Bract mosaic virus	Class 7	OL streak virus
Class2	Mild mosaic virus	Class 8	VN streak virus
Class3	Streak virus straubAcuminata Vietnam	Class 9	UM streak virus
Class4	CA streak virus	Class 10	UL streak virus
Class5	GF streak virus	Class 11	UI streak virus
Class6	IM streak virus	Class 12	UA steak virus

According to the results of experiments, we classified 12 classes into 4 groups by comparing similarities between them (see Table 1 and Table 2).

Table 2. Classification of Classes

Group	Class	Group	Class
Group 1	4,6,7,8,9,10,11	Group 3	12
Group 2	3,5	Group 4	1,2

3.2. Analysis

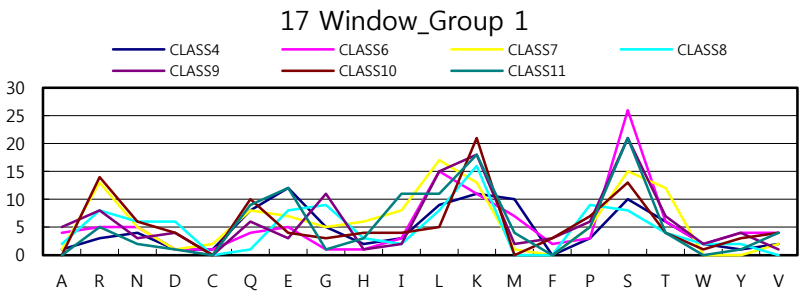


Fig. 1. 17-window group 1 amino acid.

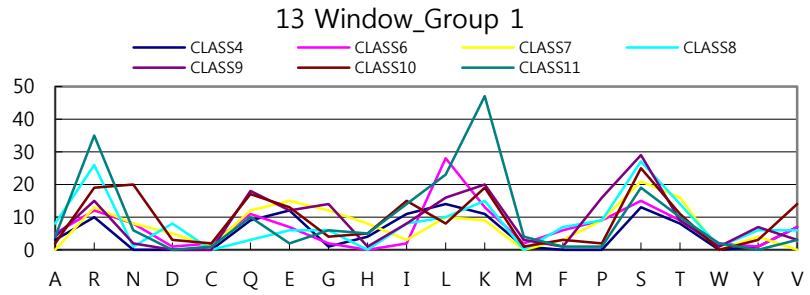


Fig. 2. 13-window group 2 amino acid.

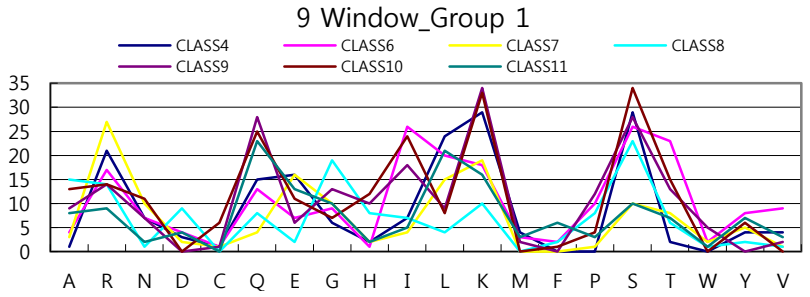


Fig. 3. 9-window group 1 amino acid.

Group 1 shows the highest percentage of amino acid 'K' and 'S'. Some of those classes also show high percentage on amino acid 'E', 'R', 'L'. Group 1 includes most of the classes we investigated, and then we inferred that these classes are one of the most common streak viruses and show the general aspect of banana viruses (see Fig. 1-Fig. 3).

Group 2 includes class 3 and 5, amino acid 'R' appeared most frequently and its disparity with other acids was evident compare to other classes. Especially, class 3 shows remarkably high percentage of amino acid 'R' compare to other classes. Other amino acids were relatively regular, and there was no notable data (see Fig. 4-Fig. 6).

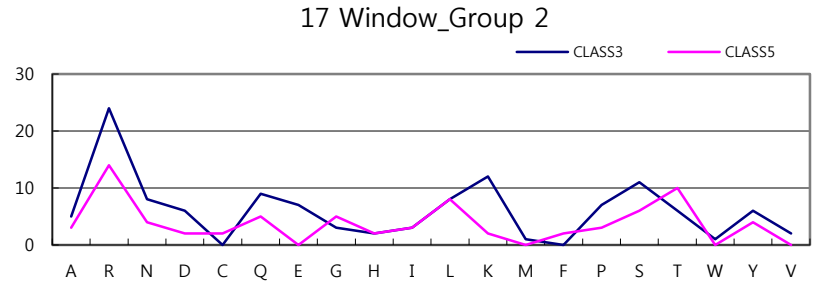


Fig. 4. 17-window group 2 amino acid.

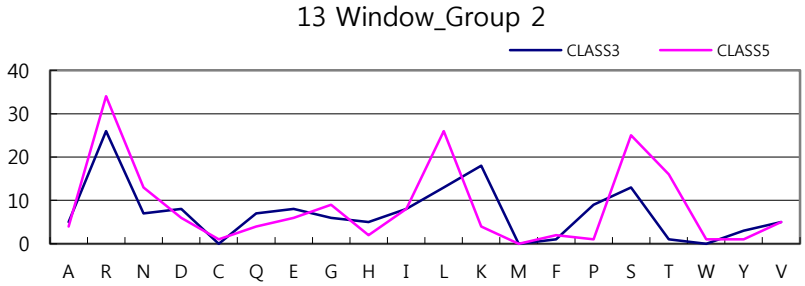


Fig. 5. 13-window group 2 amino acid.

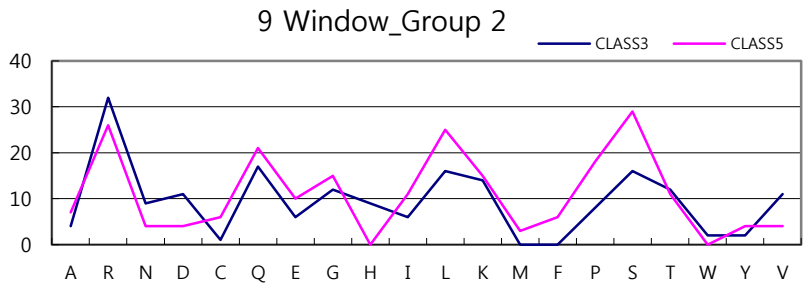


Fig. 6. 9-window Group 2 amino acid.

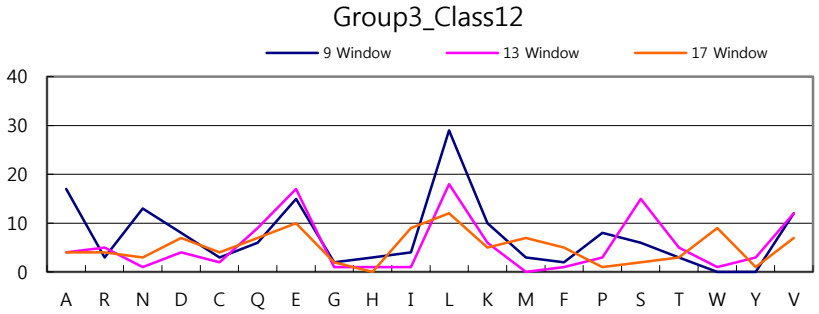


Fig. 7. Group 3 amino acid of 3 other windows.

Group 3 only includes class 12, this is because class 12 had unique features and was not perfectly consistent with other classes. It shows low percentage of amino acid 'K' and 'S', which were shown as the highest percentage in Group 1. Also, appearance of amino acid 'R' was too low, so it was not appropriate to be included in Group 2. Therefore, we classify class 12 as one single Group (see Fig. 7).

Group 4 includes class 1 and 2 (bract mosaic virus and mild mosaic virus each). Actually there is no similarity between them, but they both show distinctive feature compare to other classes and cannot be classified as same Group. They show quite different property, and this is because they are mosaic virus, while other classes are streak virus. In the graph below, reddish lines are Class 1 and blueish lines are Class 2. Three lines in one class mean 9 window, 13 window and 17 window each (see Fig. 8).

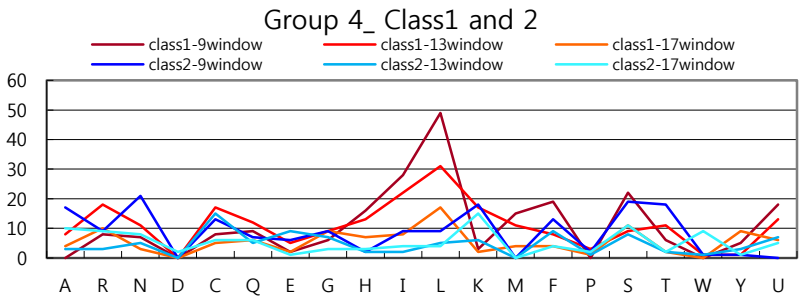


Fig. 8. Group 4 amino acid of 3 other windows.

Class 1 is the most important virus we focus on, as class 1 is speculated as the original virus that other viruses were derived from. Thus, it shows fairly general figure (deviation of values are small) that can put other classes together. We would mention this again later in conclusion.

Though the class 2 partly shows similarities with Group 1 regarding the high rate of amino acid 'K' and 'S', it shows slightly different features like considerably high percentage of amino acid 'A' and 'W', which appear rarely in other classes.

4. Rules

According to results of the experiment, we've collected approximately 7000 rules through decision tree and choose valuable data, which have considerably high frequency above 0.75. Among them, we pick rulesagain that repeatedly appear through the whole folds and have remarkable value. These rules would have to show notable features of each class, and we expect to be able to know the characteristics of each virus through these rules. These are the rules we arranged (see Table 3).

Table 3. Rule Extraction under 9 Window

Virus	Rule	Frequency	Times
Class 1	pos1=I, pos4=F	0.800	6
Class 2	pos4=A, pos5=N	0.750	7
	pos4=A, pos5=I, pos9=T	0.750	7
Class 3	pos4=H, pos7=L, pos9=R	0.750	4
Class 4	pos4=K, pos9=K	0.750	5
	pos6=S	0.750	5
Class 5	pos1=G, pos4=P, pos9=L	0.750	6
Class 6	pos2=G, pos4=I, pos5=V	0.750	8
	pos2=R, pos8=Q	0.750	7
Class 7	pos3=E, pos4=G, pos6=R	0.750	5
Class 8	pos3=G, pos4=D	0.800	6
	pos1=G, pos2=A, pos4=S	0.750	6
Class 9	pos2=G, pos4=I, pos5=I	0.750	6
	pos2=Q, pos3=T, pos4=Q	0.750	7
Class 10	pos3=S, pos4=Y, pos7=K	0.750	6
Class 11	pos4=K, pos9=Q	0.750	4
	pos1=T, pos4=E, pos5=E	0.750	4
	pos3=K, pos4=D, pos9=Q	0.750	4
	pos4=L, pos8=Y, pos9=A	0.750	4
	pos3=K, pos4=G, pos6=K	0.750	4
Class 12	pos1=A, pos4=A, po5=L	0.750	6

Table 4. Rule Extraction under 13 Window

Virus	Rule	Frequency	Times
Class 1	pos5=F, pos8=L, pos11=R	0.750	6
	pos8=C, pos13=V	0.800	6
Class 2	pos8=C, pos13=C	0.750	4
Class 3	pos8=L, pos12=R	0.750	5
Class 4	pos5=L, pos6=K, pos8=T	0.750	4
	pos4=S, pos5=V, pos8=T	0.750	4
	pos1=I, pos8=I, pos13=S	0.750	4
Class 5	pos5=I, pos8=R	0.750	6
	pos1=R, pos8=L, pos11=T	0.750	6
	pos1=S, pos8=S, pos9=N	0.800	6
Class 6	pos8=S, pos9=L, pos11=P	0.750	6
Class 7	pos8=L, pos9=E, pos11=P	0.750	5
Class 8	pos5=A, pos8=T, pos9=P	0.800	5
	pos8=L, pos12=S	0.750	5
Class 9	pos6=T, pos7=Q	0.800	6
Class 10	pos4=Q, pos5=V, pos8=T	0.750	5
	pos6=V, pos8=L, pos11=I	0.750	5
	pos1=N, pos8=S, pos12=E	0.750	5
Class 11	pos8=L, pos11=S, pos12=R	0.750	5
Class 12	pos1=S, pos8=S, pos9=Q	0.750	5

We've found that each window has specific position frequently appears in rules. Position 4 in 9 window, position 8 in 13 Window, and position 15 in 17 Window were found in most rules of each Window. These rules were selected and organized from those that were more probable, and the position between the windows turned out to be irrelevant, because the designated range of sequence were all different.

Frequencies being all high, we focused on the number of appearances, and the rules mentioned above are ones that are the modes, and are rules distinct from other classes. Through these rules, we are looking forward to detect not only features of each banana viruses, but also their general genetic tendency (Table 4 and Table 5).

Table 5. Rule Extraction under 17 Window

Virus	Rule	Frequency	Times
Class 1	pos1=Y, pos15=Y	0.750	3
Class 2	pos1=R, pos8=A, pos15=S	0.750	5
	pos8=Q, pos15=W	0.750	5
Class 3	pos4=T, pos15=D, pos14=S	0.750	3
	pos2=R, pos15=E	0.750	3
Class 4	pos2=E, pos15=M	0.750	6
Class 5	pos1=G, pos12=T, pos15=L	0.750	3
Class 6	pos2=S, pos12=S, pos15=M	0.750	4
Class 7	pos8=G, pos11=E, pos15=K	0.750	4
Class 8	pos15=D, pos16=N	0.750	6
Class 9	pos4=S, pos15=P	0.750	5
	pos9=G, pos15=F	0.750	5
	pos8=L, pos9=L, pos15=A	0.750	5
Class 10	pos4=P, pos8=Q, pos15=S	0.800	4
Class 11	pos11=S, pos12=S, pos15=K	0.750	3
Class 12	pos15=D, pos16=W	0.750	6

In the above are the rules from 13 window and 17 window.

5. Conclusion

This study is a synthesis of banana research, and our purpose is to seek ways to cure banana viruses by analyzing its amino acid and sequences. We have found that there are some amino acids that appear frequently in most classes (amino acid 'R','K' and 'S') and some positions selected as rules repeatedly, and these are the general but unique features that banana viruses have. According to the summary, bract mosaic virus (class 1) seems to contain most of sequences. It takes most of its sequences and absorbs sequences of other viruses also. We suggest that bract mosaic virus has strong identity and other banana viruses were derived from it. Also, there were lots of rules and high rates of errors, which mean that those viruses are individually distinct entities, and each virus has their own characteristics. As we proceed with experiment, 17 window was usually be the standard of our analysis, and we have found that 17 Window shows similarities between classes clearly and more accurately. As a result of analysis of amino acid and sequences, viruses were divided into 4 groups. We arranged amino acid in frequency-descending order, and classified 12 classes according to their characteristics. Group 4, which includes bract mosaic virus and mild mosaic virus, shows different features compare to other streak viruses which included in Group 1, 2, and 3. As the charts above suggest, each viruses had their own rules and features that cannot be bound into one single category, and we analyzed those viruses one by one. We are looking forward to find ways to cure banana viruses through this research. Inconsistency of the results and difficulty of finding similarities between Windows was limit of our research. Some classes (especially class 2) shows inconsistent results like amino acids which appear frequently in 17 Window do not in 9 Window. Because of these unstable outcomes, we have trouble in drawing conclusion. Also, as we have collected too broad range of raw data due to too many classes, shifting out valuable data was patience-needed progress.We have tried hard to form amost efficientalgorithm thatcan distinguishthebananavirus, and finally obtaineddatain consideration ofthe similarity betweenDNAsequences of thebananaviruses byusingdecision tree.Among the obtainedlarge amounts of data, extracting and organizing meaningfulvalues, and classifying viruses into group according

to the data were our tasks.This paper is a unique research which deals with almost all kinds of banana streak viruses. In terms of the fact that we have preceded our research in data analytic method with decision tree algorithm, it shows differentiation from other researches.

References

[1] Iskra-Caruana, M. L., Chabannes, M., Duroy, P. O., & Muller, E. (2014). A possible scenario for the evolution of Banana streak virus in banana. *Virus Research, 186*, 155-162.

[2] Dahal, J., & Lockhart, B. E. L. (1998). Status of banana streak disease in Africa: Problems and future research needs. *Integrated Pest Management Reviews, 3(2)*, 85-97.

[3] Hull, R., & Glyn H. (1998). Cloning and sequence analysis of Banana streak virus DNA. *Virus Genes*, *17(3)*, 271-78.

[4] Glyn, H., Ganesh, D., Thottappilly, G., & Hull, R. (1999). Detection of Episomal banana streak bad navirus by IC-PCR. *Journal of Virological Methods, 79(1),* 1-8.

[5] Baranwal, V. K., Sharma, S. K., Khurana, D., Verma, R. (2014). Sequence analysis of shorter than genome length episomal Banana streak OL virus like sequences isolated from banana in India. *Virus Genes, 48(1)*, 120-127.

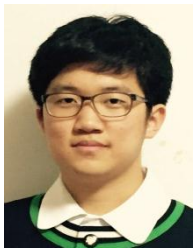
[6] Geering, A. D. W., McMichael, L. A., Dietzgen, R. G., & Thomas, J. E. (2000). Genetic diversity among Banana streak virus isolates from Australia. *Phytopathology*, *90(8)*, 921-927.

[7] Go, E., Lee, S., & Yoon, T. (2014). Analysis of Ebola virus with decision tree and Apriori algorithm. *International Journal of Machine Learning and Computing, 4(6)*, 543-546.

[8] Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

[9] U.S. National Library of Medicine. (Oct. 2014). *National Center for Biotechnology Information*.

Seohee Lee was born in Republic of Korea, in 1997. She is a student of Hankuk Academy of Foreign Studies in Yong-in, Korea. She is majored in chemistry and biology. She has been recently interested in bio-informatics and started her first research in 2014. Her main subject is analyzing DNA sequences of viruses, and she is looking forward to proceed study in other viruses.



Wonil Roh was born in Yong-in, South Korea in 1997. He is currently studying at Hankuk Academy of Foreign Studies in Yongin, Gyeonggi, South Korea. He wants to major plant biology and biology education in the future.

Taeseon Yoon was born in Seoul, Korea, in 1972. He got a Ph.D. 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, as a lecturer and Ansan University, as an adjunct professor. Since December 2004, he has been with the Hankuk Academy of Foreign Studies, where he is a computer science and statistics teacher.