Analysis of Soil-Borne Wheat Mosaic Virus and Soil-Borne Cereal Mosaic Virus Using Datamining

Bogyung Kim, Jiin Jeong, and Taeseon Yoon
Department of International, Hankuk Academy of Foreign Studies, Mohyeon-myeon, Yongin-si, South Korea, 17035
Email: {juliakimm98, luckyjiin98}@gmail.com, tsyoon@hafs.hs.kr

***Abstract*—Plant viruses are viruses that affect the proper development of plants. The research for plant viruses, first initiated by the scientist A. Mayer in 1886, grew to offer various explanations for their transmission, including vector transmission, generation-to-generation transmission from seed and pollen, and the rare human-plant transmission. Among the many ways which the plant viruses can spread through, a small number of plant viruses were found to become transmitted through the soil. Such two examples of soil-borne viruses that infect plants through soil are soil-borne wheat mosaic virus (SBWMV) and soil-borne cereal mosaic virus (SBCMV). In this paper, we compared the protein sequences encoded in the RNA genomes of soil-borne cereal mosaic viruses and soil-borne wheat mosaic viruses, particularly focusing on replication protein, coat protein, cysteine rich protein, and 84-kDa protein. Furthermore, through examination of the position and frequency of each amino acid, we analyzed the similarities and differences between the two soil-borne viruses.**

***Index Terms*—Apriori algorithm, decision tree, proteins, soil-borne cereal mosaic virus, soil-borne wheat mosaic virus**

I. Introduction

Soil-borne wheat mosaic virus (SBWMV) and soil-borne cereal mosaic virus (SBCMV) are two types of plant viruses that infect plants, impeding their proper growth and development. A breakdown of the names of the soil-borne wheat mosaic virus and soil-borne cereal mosaic virus demonstrates the properties each virus contains. The two viruses are soil-borne viruses, transmitted through the soil, with or without vectors, depending on circumstances. Vectors for spread of the soil-borne viruses are either fungal or plant-parasitic nematodes [1]. It is possible for viruses to travel abiotically from the infected roots [2]. Usually, infection by soil-borne viruses takes place in the tips of young and metabolically active roots, rather than the larger and older roots [3].

Soil-borne viruses cause a serious damage to crops, affecting the economically important ones such as wheat, barley, potato, and fruit crops. As can be seen from their names, SBWMV and SBCMV infect cereal grains. A cereal is a type of grass that is grown for an agricultural purpose to produce edible grain. Wheat is a cereal grain, being one of the top most-produced types of cereal in the world. The relation between wheat and cereal makes it even more complex to distinguish between SBWMV and SBCMV. Also, because of the complicated plant-vector-virus interaction and the soil interfering with the research, there is still an overall lack of understanding of soil-borne viruses [4]. Furthermore, the viruses' prolonged existence and the absence of effectual strategy make it a crucial task to experiment and research in-depth about such viruses.

SBWMV and SBCMV have several common characteristics. First, plant viruses have 73 genera and 49 families, and soil-borne plant viruses are grouped into fifteen named and two unnamed genera. Out of these, both SBWMV and SBCMV belong to the genus Furovirus, a genus of viruses in the family Virgaviridae. Structure-wise, furovirus is non-enveloped and rod shaped, having helical symmetry. The virion consists of two portions, one long and one short, which are each 140-160 nm and 260-300nm long and 20 nm wide [5]. The name Virgaviridae derives from the Latin word virga (rod), which accurately describes the rod-shaped structure of viruses in the Virgaviridae family. Virgaviridae comprises of positive single stranded RNA viruses and uses plants as its natural host [6]. Another similarity is that both viruses are transmitted by the vector plasmodiophorid Polymyxa graminis. These parasites were traditionally regarded as fungi, but were recently found to resemble protists. Polymyxa graminis is itself non-pathogenic, but it carries various plant diseases, causing serious diseases in and reducing the yield of cereal crops. Lastly, the two viruses are types of mosaic viruses, which cause a speckled and yellowing look in the leaves of plants. SBWMV, in particular, is the cause of green and yellow mosaic.

On the other hand, the two viruses differ in occurrence when it comes to time and place. The common locations for these two viruses differ as soil-borne wheat mosaic virus is mainly found in North America and Chinese wheat mosaic virus in Asia, whereas soil-borne cereal mosaic virus most frequently appears in Europe [7]. Also, while there is some controversy with some claiming SBWMV infection to take place in spring, SBWMV is primarily known as a serious

Manuscript received August 15, 2016. revised January 12, 2017.

doi: 10.18178/ijpmbs.6.1.11-15

winter wheat disease. Stunting and mosaic symptoms mostly occur in early spring for SBCMV [3].

This paper intends to compare the protein sequences encoded in the RNA genomes of SBWMV and SBCMV, particularly focusing on replication protein, coat protein, cysteine rich protein, and 84-kDa protein. To do this, we used two methods of datamining, decision trees and apriori algorithm, which will be explained in detail in the following section.

II. Materials and Methods

A. Proteins

The two viruses consist of two particles, RNA 1 and RNA 2, where there are proteins encoded. Replication in plant viruses observes the positive stranded RNA virus replication model, using positive stranded RNA virus transcription as its method of transition. The replication takes place in the cell's cytoplasm. In soil-borne wheat mosaic virus, the replication protein is in RNA 1, the longer strand of the two particles. Two proteins in RNA 1, measuring 150-kDA and 209-kDa, are responsible for virus replication. The coat protein or the capsid proteins package viral genetic materials inside capsids and moderate binding to host cells. In SBWMV, the coat protein is located in RNA 2, the shorter particle, and is called the 19-kDa coat or Capsid Protein (CP). Cysteine rich protein suppresses the post transcriptional gene silencing, which is a mechanism that assists the host in resisting a virus. It is a 19-kDa cysteine-rich protein and is expressed through a subgenomic mRNA. 84-kDa protein serves the function of virus transmission. It is related with another protein, 37-kDa, which allows viral cell-to-cell movement. This is also expressed through a subgenomic mRNA [8]. We analyzed these four proteins for we found them crucial in the spread of the soil-borne wheat mosaic virus and the soil-borne cereal mosaic virus.

B. Decision Tree

A decision tree is a visual model that uses a tree-like structure to graphically illustrate decisions and their possible consequences [9]. The tree consists of three types of nodes and two types of branches. The three nodes, which are root decision node, chance node, and terminal node, are each assigned class labels. A root decision node, also known as choice node, is commonly represented by a square and shows the available options to the decision maker. It has no incoming lines and can have zero to multiple outgoing lines. A chance node and a terminal node are found at the end of the initial branches. In most cases, a chance node is uses the shape circle and is called an uncertainty node because it suggests the possible outcomes and their probabilities from 0 to 1.0. In contrast, a terminal node shows a fixed value [10]. The branches drawn from left to right and the path from root to leaf each demonstrates the possible actions and the classification rules. As one method of displaying algorithm, the decision tree analyzes decisions, calculates an outcome's probability, and finds out the most optimal strategy available.

Since it is a graphic and hierarchical model, a decision tree is easy to understand, being frequently used in operations management or research to determine a strategy that is highly likely to accomplish a goal. However, in cases of where there are too little data or too many outcomes, the calculations can turn complex. This paper focuses on the decision tree composed of windows 9, 13, and 17, and experimented with replication protein, coat protein, cysteine rich protein, and 84-kDa protein. The rules of the decision were impossible to find when the subjects bore too much similarity to one another.

C. Apriori Algorithm

Apriori is an algorithm that uses the simplest set class theory to extract the frequency of the elements and association rules from large data sets [11]. It functions on databases containing transactions, or sets of items. This data mining method is used to identify the rules which satisfy a minimum support threshold and a minimum confidence threshold. In short, the apriori algorithm finds all frequent itemsets that have presence in database greater than or equal to the minimum support threshold. From those, the apriori generates strong association rules [12]. Association rules help predict the frequency of an item or item set in a certain set of transaction by using the occurrence of other items in the transaction.

The main observation of apriori principle is that for a frequent itemset, all its subsets will occur frequently as well. This holds true as illustrated by the following property of the support measure, which shows that the support of an itemset never exceeds the support of its subsets:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

Apriori uses a "bottom up" approach to stretch frequent subsets and test a group of candidates against the data until there can no longer be successful extensions or candidate generation. Just like the decision tree, the experiment was carried out with windows 9, 13, and 17. The results prove the frequency of the positions certain types of amino acids appeared.

III. Experiment

A. Decision Tree

TABLE I. Rule Extraction of RP under 9 Window

Virus	Rule	Frequency
SBWMV	Pos1=I	0.750
	Pos5=E	0.800
SBCMV	Pos1=H	0.700
	Pos4=K	0.700

TABLE II. Rule Extraction of RP under 13 Window

Virus	Rule	Frequency
SBWMV	Pos2=R	0.833
	Pos2=D	0.800
SBCMV	Pos2=P	0.833
	Pos1=A	0.750
	Pos4=G	0.750

The results from Table I show that both viruses have highest amino acid frequency at position 1 under window 9, while their frequencies are different from 750 to 700. Table II gives the clue that position 2 presents significant role in both viruses, because position 2 shows in the same highest frequency in both viruses under window 13. Table III clearly shows that rules were only extracted in position 8. However, in the case of SBWMV, pos8 = A was the most frequent, while in the case of SBCMV, pos8 = E was the most frequent. Also, when looking at the graph as a whole, the frequency of amino acid extracted at position 8 of SBWMV is higher than that of SBCMV. No other relationships are proven.

TABLE III. RULE EXTRACTION OF CP UNDER 9 WINDOW

Virus	Rule	Frequency
SBWMV	Pos1=A	0.800
	Pos1=P	0.800
	Pos1=I	0.800
SBCMV	Pos1=R	0.750
	Pos1=N	0.750

According to Table IV, the only rule extracted from CP (Coat Protein) was from the position 1 under window 9. So, we assume that position 1 will be the crucial factor which makes SBWMV and SBCMV different from each other. Moreover, the SBWMV showed higher amino acid frequency compared to SBCMV.

No rules had been extracted from Cysteine Rich Protein of both viruses.

TABLE IV. RULE EXTRACTION OF 84KDA PROTEIN UNDER 9 WINDOW

Virus	Rule	Frequency
SBWMV	Pos5=N	0.818
	Pos7=H	0.857
	Pos1=K	0.750
SBCMV	Pos5=N	0.800
	Pos3=D	0.750

TABLE V. RULE EXTRACTION OF 84KDA PROTEIN UNDER 13 WINDOW

Virus	Rule	Frequency
SBWMV	Pos3=Y	0.750
SBCMV	Pos3=N	0.800

TABLE VI. RULE EXTRACTION OF 84KDA PROTEIN UNDER 17 WINDOW

Virus	Rule	Frequency
SBWMV	Not extracted	
SBCMV	Not extracted	

Table V conveys that there are similarities in pos5=N between 84-kDa protein of SBWMV and SBCMV. Although their frequency is slightly different from 0.818 to 0.800, this sharing of the amino acid in same position might deeply contribute to the sameness of two viruses. Also, according to Table VI, like the SBWMV, in SBCMV, position 3 occurs most frequently. This implies position 3 plays important roles in both viruses. Table VIII shows that there is no particular relationship between 84kDa protein of SBWMV and SBCMV under window 17.

B. *Apriori Algorithm*

The results from Apriori Algorithm showed following patterns.

1. amino1=T 23
2. amino3=T 18
3. amino2=T 17
4. amino5=L 17
5. amino4=T 16
6. amino2=S 15
7. amino3=K 14
8. amino4=L 14

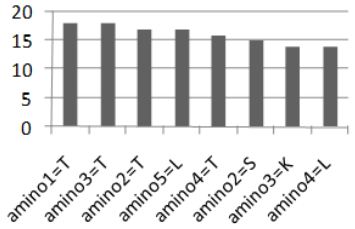

Figure 1. 9-window RP of soil borne wheat mosaic virus amino sequence

Fig. 1 is the result presented by analyzing RP(Replication Protein) of Soil Borne Wheat Mosaic Virus under 9 windows. Other proteins analyzed had similar patterns. The most frequent rules of each protein of both viruses were selected and provided. In the graph, the types of amino acid are arranged by the frequency of them in order to reveal the similarities in protein structures between Soil Borne Wheat Mosaic Virus and Soil Borne Cereal Mosaic Virus. The results are shown in the below.

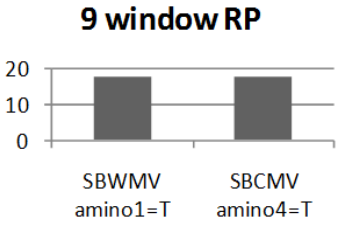

Figure 2. 9-window RP amino sequence of SBWMV and SBCMV

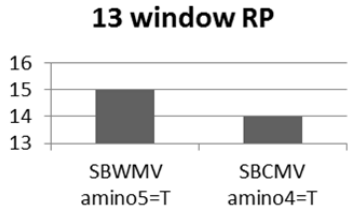

Figure 3. 13-window RP amino sequence of SBWMV and SBCMV

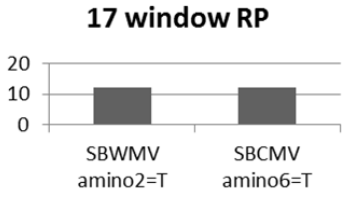

Figure 4. 17-window RP amino sequence of SBWMV and SBCMV

Each of Fig. 2, Fig. 3, and Fig. 4 shows the analysis of RP(Replication Protein) of SBWMV (Soil-Borne Wheat Mosaic Virus) and SBCMV (Soil-Borne Cereal Mosaic Virus) under windows 9, 13, and 17. 9-window: Comparison between SBWMV amino1 Thiamine and SBCMV amino4 Thiamine. 13-window: SBWMV amino5 Thiamine and SBCMV amino4 Thiamine. 17-window: SBWMV amino2 Thiamine and SBCMV amino4 Thiamine. In all 3 windows, both viruses showed the highest frequency in Thiamine. Especially, in 9-window and 17-window analysis, SBWMV and SBCMV presented same frequency in Thiamine. They not only shared the highest amino acid but they also had same frequency of it. This clearly proves the probability of similarity between these two viruses.

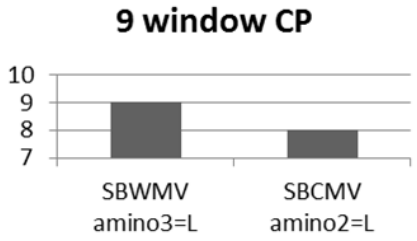

Figure 5. 9-window CP amino sequence of SBWMV and SBCMV

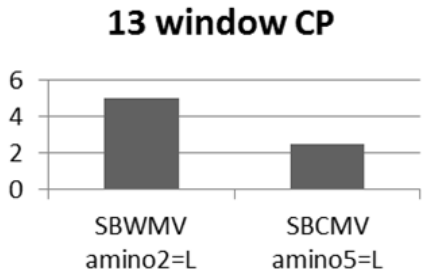

Figure 6. 13-window CP amino sequence of SBWMV and SBCMV

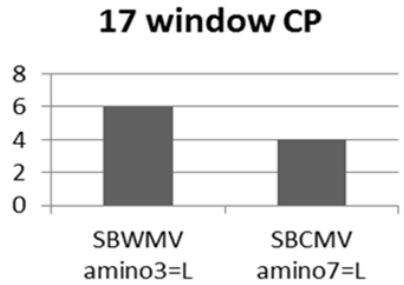

Figure 7. 17-window CP amino sequence of SBWMV and SBCMV

Fig. 5, Fig. 6, and Fig. 7 are the analysis of CP of both viruses under windows 9, 13, and 17. 9-window: Comparison between SBWMV amino3 Leucine and SBCMV amino2 Leucine. 13-window: SBWMV amino2 Leucine and SBCMV amino5 Leucine. 17-window: SBWMV amino3 Leucine and SBCMV amino7 Leucine. The result shows that SBWMV has more Leucine in CP than that of SBCMV.

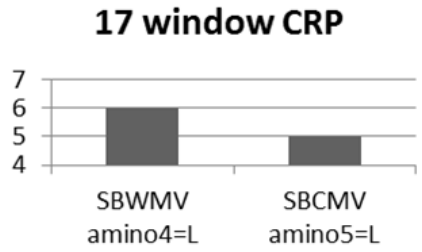

Figure 8. 17-window CRP amino sequence of SBWMV and SBCMV

Fig. 8 is the analysis of CRP (Cysteine Rich Protein) of both viruses under 17 windows. Under 9 and 13 windows, both viruses had no common amino acids to compare with, so comparison was possible only in 17 windows. 17-window: Comparison between SBWMV amino4 Leucine and SBCMV amino5 Leucine. In this case, Leucine was the sharing amino acid having highest frequency, and SBWMV had higher Leucine.

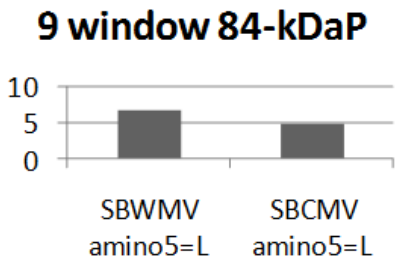

Figure 9. 9-window 84-kDa Protein amino sequence of SBWMV and SBCMV

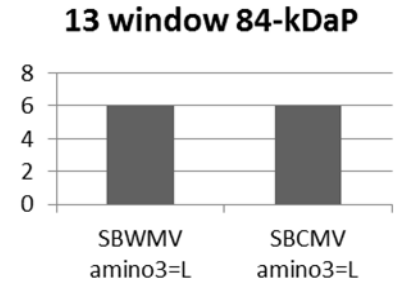

Figure 10. 13-window 84-kDa Protein amino sequence of SBWMV and SBCMV

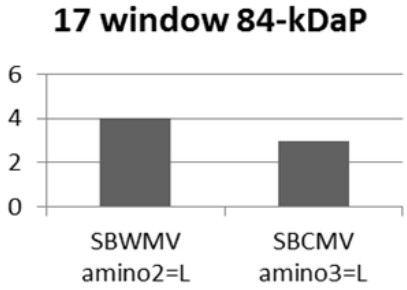

Figure 11. 17-window 84-kDa Protein amino sequence of SBWMV and SBCMV

Fig. 9, Fig. 10, and Fig. 11 are the analysis of 84-kDa protein of SBWMV and SBCMV under windows 9, 13, and 17-window: Comparison between SBWMV amino5

Leucine and SBCMV amino5 Leucine. 13-window: SBWMV amino3 Leucine and SBCMV amino3 Leucine. 17-window: SBWMV amino2 Leucine and SBCMV amino3 Leucine. As the result shows, both viruses had highest Leucine at same position except in window 17. Thus, position 3 and position 5 should be important to the similarities between the two viruses. The analyzed result also reveals that generally Leucine in 84-kDa protein of SBWMV is higher than that of SBCMV.

IV. Conclusion

The analysis by Apriori algorithm showed that the higher the windows become, the more various the types of DNAs were. After observing the results from Decision Tree, it was able to find similarities between soil-borne wheat mosaic virus and soil-borne cereal mosaic virus. Although some windows failed to extract rules, it was definite that they share some common features for RP, CP, CRP, and 84-kDa protein. Since these two viruses are controversial by their ambiguous boundary, this research would be helpful providing some properties they possess similarly.

As research on soil-borne viruses is still inadequate, we believe that this paper will help categorize the proper relationship and treatment for soil-borne virus as well as raise awareness in the importance of such investigation. For further research, we want to examine the acidity and proportion of nitrogen in the soils of each continent to analyze why there is a difference in the distribution of soil-borne cereal mosaic virus and soil-borne wheat mosaic virus.

References

[1] K. M. Smith, *Plant Viruses*, Springer Science & Business Media, 2012.
[2] A. G. Roberts, *Plant Viruses: Soil-borne*, eLs, 2014.
[3] C. Hiruki and D. S. Teakle, "Soil-Borne Viruses of Plants," in *Current Topics in Vector Research*, K. F. Harris, Ed., New York: Springer, 1987, pp. 177-215.
[4] D. Perovic, K. Kanyuka, and F. Ordon, "Disease resistance. Soil-borne cereal mosaic virus (SBCMV) resistance in bread wheat," *Mas Wheat.*
[5] SIB Swiss Institute of Bioinformatics, Furovirus, *ViralZone*, Web, 2012.
[6] SIB Swiss Institute of Bioinformatics, Virgaviridae, *ViralZone*, Web, 2012.
[7] T. Kühne, "Soil-borne viruses affecting cereals: Known for long but still a threat," in *Virus Research*, J. Michael Thresh, A. Fereres, Roger, A. C. Jones, and P. Lava , Ed., Elsevier, 2009, pp. 174-183.
[8] L. Cadle-Davidson and S. M. Gray, "Soil-borne wheat mosaic virus," *The Plant Health Instructor*, 2006.
[9] C. Kingsford and S. L. Salzberg, "What are decision trees? " *Nature Biotechnology*, vol. 26, pp. 1011-1013, 2008.
[10] B. Brumen, *Introduction to Decision Tree*, Maribor: U of Maribor, Institut of Informatics, 2002.
[11] Z. Shi, *Advanced Artificial Intelligence*, Singapore: World Scientific, 2011.
[12] O. G. Salem, "Apriori Algorithm," *Code Project*, 2012.

Bogyung Kim was born in Busan, Republic of Korea in 1998. She is currently enrolled in the Department of International in Hankuk Academy of Foreign Studies. She co-authored "Analysis of Ebolavirus and Marburgvirus using Datamining", which was published in the 9th International Conference on Bioinformatics and Biomedical Engineering (Shanghai, China: CRC, 2015). Her research focuses on bioinformatics, computer science, and statistics.

Jiin Jeong was born in Iksan, Republic of Korea in 1998. She is currently enrolled in the Department of International in Hankuk Academy of Foreign Studies. She co-authored "Analysis of Ebolavirus and Marburgvirus using Datamining" which was published in the 9th International Conference on Bioinformatics and Biomedical Engineering (Shanghai, China: CRC, 2015). Her research focuses on bioinformatics, biomedical engineering, and computer science.

Taeseon Yoon was born in Seoul, Republic of Korea in 1972. He received his Ph. D. degree in computer education from Korea University in 2003. Since December 2004, he has been with Hankuk Academy of Foreign Studies where he is a Computer Science and Statistics teacher.