ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 320-324
doi: 10.18178/wcse.2016.06.049

Analysis of the Genomes of Chikungunya Virus and Dengue Virus Using Decision Tree, Apriori Algorithm, and Support Vector Machine

Donghyun Lee [1+] and Taeson Yoon [2]

[1] International Studies: Hankuk Academy of Foreign Studies, Yongin, Republic of Korea
[2] Computer Science: Korea University, Seoul, Republic of Korea

Abstract. Chikungunya Virus, which has no vaccines or medicines, is introduced into human population by mosquito bites. It is mostly occurred in Africa, Asia, Europe, and the Indian and Pacific Oceans. Dengue Virus, leading cause of death in the tropics and subtropics, is also transmitted by mosquitoes: 400 million infections every year. Chikungunya Virus and Dengue Virus share their clinical presentations such as joint pain, severe fever, and rash. There is no vaccine to prevent either Chikungunya Virus or Dengue Virus. In this paper, I analysed the DNA sequence of CHIKV and DENV in order to investigate the difference between them. Furthermore, I looked for the amino acid frequencies and found the genetic similarities and correlation between CHIKV and DENV.

Keywords: Chikungunya virus, dengue virus, decision tree, data mining, Apriori algorithm, support vector machine.

1. Introduction

Chikungunya and Dengue are both co-circulating vetor-borne diseases with substantial overlap in clinical presentations. Chikungunya virus (CHIKV) originated in 1953 in southern Tanzania [1]. CHIKV is a mosquito-borne alpha virus in the family *Togaviridae*. CHIKV infection causes a very sharp onset of arthralgia, severe fever and eventually the conspicuous rash. Dengue virus (DENV) is mosquito-borne virus in the family *Flaviviridae,* causing 400 million infections on yearly basis [2]. Chikunguya and Dengue share substantial clinical presentations. The classical manifestations of these two diseases have substantial overlap both diseases having fever, headache, myalgia, and rash. Looking at the studies comparing the presentations of Dengue and Chikungunya- shorter duration of fever, connjunctivitis, acute arthiritis, myalgia, artharalgia and rash were more prominent in chikungunya; whilst leukopenia, neutropenia, thrombocytopenia and abdominal patin were more prominent in Dengue cases. Although each has own characteristic symptoms, they cannot be differentiated by symptoms only. It is important to differentiate between Chikungunya and Dengue during first presentation as a clinical treatment. Although there have been active researches conducted by private and public researchers, there are no World Health Organisations pre-qualified human vaccines nor antiviral treatment against both CHIKV and DENV [3].

2. Materials

2.1. Alpha virus

Alpha virus, which can be divided into Old World Viruses and New World Viruses, has evolved distinct ways of interacting with their hosts and tropism. There are 29 different types of alpha viruses that cause dis-

[+] Corresponding author. Tel.: +821031362912.
E-mail address: donghyun919@hafs.hs.kr.

eases in human and other mammals. These species of arboviruses are classified into 7 antigen complexes: Barmah Forest (BF), Eastern equine encephalitis (EEE), Middelburg (MID), Ndumu (NDU), Semliki Forest (SF), Venezuelan equine encephalitis (VEE), and Western equine encephalitis (WEE). Chikunguya virus is classified as Smliki Forest (SF) group of Old World Alphaviruses [4].

2.2. Chikunguya virus

Chikungunya virus is a small, spherical, enveloped, positive-strand RNA virus. In 2006, the complete sequence of Chikungunya, isolated from Reunion Island was made available at National Centre for Biotechnology information [5]. CHIKV genome encodes for two proteins: the structural polyprotein consisting of five proteins (Capsid, E1, E2, E3, and 6K), and the non structural polyprotein consisting of four proteins (nsP1, nsP2, nsP3 and nsP4). The 5' end of the RNA molecule is capped with a 7-methylguanisine whilst the 3' end is poly-adenyated. A subgeneric positive strand RNA known as 26 SRNA is transcribed from a negative-stranded RNA intermediate which severs as the mRNA for the synthesis of the viral structural proteins. Alpha viruses have conserved domains that play an important role in the regulation of viral RNA synthesis. These domains are found at the 5' and 3' ends and intergenic region. The E1 and E2 glycoproteins are expected to form heterodimers, and associate as trimeric spikes on the viral surface that cover surface evenly. The envelope glycoproteins play a role in attachment to cells. Viral located on the surface of the cell membrane enters the host cells by endocytosis and fusion of the viral envelope. The uncoating of the virions occurs in the cell cytoplasm. Replication is not restricted to a particular tissue or organ of the host so the virus replication occurs in various organs. The insect host initiates the virus replication, and the genome replication is done in the cytoplasm.

2.3. Dengue virus

Dengue Virus (DENV) is a single-stranded RNA positive-strand virus of the family Flaviviridae, genus Flavivirus. In 2004, the complete sequence of Dengue, isolated from Buenos Aires, Argentina, was made available at NCBI [6]. DENV is a 50 nanometre virus enveloped with a lipid membrane. There are 180 identical copies of the envelope proteins attached to the surface of the viral membrane by a short transmembrane segment. The virus has a genome about 11,000 bases that encodes a single large polyprotein that is cleaved into several structural and non structural mature peptides. The polyprotein is divided into three structural proteins (C, prM, E), seven nonstructural proteins (NS1, NS2a, NS2b, NS3, NS4a, NS4b, NS5), and short non coding regions on both the 5' and 3' ends. The structural proteins are the capsid (C) proteins, the envelope (E) glycoprotein and the Membrane (M) protein, itself derived by purine-mediated cleavage from a prM precursor. The E glycoprotein is responsible for virion attachment to receptor and fusion of the virus envelope with the target cell membrane and bears the virus neutralisation epitopes. Only one other viral protein, NS1, has been associated with a role in protective immunity. NS3 is a protease and a helicase, whereas NS5 in the RNA polymerase in charge of viral RNA replication. DENV uncoats intracellularly via a specific process. In the infectious form of the virus, the envelope protein lays flat on the surface of the virus, forming a smooth coat with icosahedral symmetry. However, when the virus is carried into the cell and into lysosomes, the acidic environment causes the protein to snap into different shape, assembling into trimeric spike. Several hydrophobic amino acids at the top of this spike insert into the lysosomal membrane and cause the virus membrane to fuse with lysosomes. This release the RNA into the cell and infection starts.

3. Methods

3.1. Decision tree

Decision tree is an analysis method commonly used in data mining, in order to create a model that predicts the value of a target variable based on several input variables. This method uses a decision tree as a predictive model. I extracted the genome sequences of DENV and CHIKV from the National Centre for Biotechnology Information (NCBI). For the decision tree, I used the sequences for a 10-fold cross validation experiment, which is a model validation technique for assessing how the result of a statistical analysis will generalise to the independent set. I extracted data with very high frequency rates; only with frequency rates higher than 0.83 were chosen.

3.2. Apriori algorithm

Apriori algorithm is the classical form of algorithm for finding and extending frequent items in a transactional database. This algorithm accentuates the general trends of the database and reveals the association rules. Using bottom up approach, frequent subsets are extended one at a time, which is known as candidate generation, and each group of candidates is tested against the data until no further extensions are found [7]. By Breadth-first search and a Hash tree structure, candidate item sets of length K are generated from length k-1, and then the candidates with infrequent sub pattern are removed. According to downward closure lemma, the candidate set comprises every frequent k-length item set. Then, transaction database is scanned to reveal the frequent item sets amongst the candidates. In this experiment, I defined class 1 as DENV and class 2 as CHIKV. I divided the amino acid sequences of DENV and CHIKV into 9, 13, 17 windows and analysed the frequent sequence pattern respectively.

3.3. Support vector machine

Support Vector Machine (SVM) is a supervised learning model which is kernel based techniques such as Gaussian processes, Bayes point machines, and kernel principal component analysis. Since SVM not only shows similar accuracy with Artificial Neural Network (ANN), but also complement weakness of ANN such as local optimisation [8]. There are four functions of SVM used in this research: Polynomial 1, Polynomial 2, Normal, RBF, and Sigmoid. Polynomial Kernel is one of Kernel functions commonly used along with SVM and other models. It represents the similarity of training samples in a feature space over polynomials of the original variables. Polynomial 1 is a SVM function related to linear function, whilst Polynomial 2 is a SVM function related to non-linear function. RBF Kernel, which is also known as Gaussian radial basis function kernel, is a widely used SVM classification. Since SVM has difficulties in scaling a large number of training samples and features in the input space, several approximations to the RBF kernel have been devised. In this research, 10 fold cross validation is used.

4. Results

Table 1: SVM 9 Window

Function	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Average
Poly 1	55.94	56.25	58.44	56.56	57.03	57.57	57.51	59.06	61.56	60.00	57.992
Poly 2	56.41	48.75	53.91	52.34	45.78	56.41	43.75	58.75	52.97	52.81	46.907
Normal	44.06	50.00	48.44	48.11	50.47	47.50	47.03	47.81	45.47	49.06	47.795
RBF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Sigmoid	55.94	50.00	52.19	50.47	52.50	52.99	52.19	54.53	50.54	51.56	52.269

Table 2: SVM 13 Window

Function	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Average
Poly 1	63.75	67.71	62.29	67.71	67.92	63.75	61.04	59.17	61.88	64.79	64.001
Poly 2	57.50	52.29	58.12	51.88	57.50	49.79	55.62	56.04	59.17	61.67	55.958
Normal	49.17	42.92	46.04	46.46	45.62	45.83	49.38	50.83	47.92	48.75	47.292
RBF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Sigmoid	50.83	57.08	53.96	53.54	54.38	54.17	50.62	51.04	52.08	51.25	52.895

4.1. Support vector machine

Table 3: SVM 17 Window

Function	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Average
Poly 1	63.14	65.71	65.71	65.43	66.57	64.86	66.29	67.86	65.14	63.14	65.384
Poly 2	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Normal	48.57	49.14	47.14	42.00	45.43	46.57	48.86	45.14	44.57	51.43	46.869
RBF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Sigmoid	53.14	50.86	52.80	58.00	54.57	53.43	54.14	54.86	55.43	51.43	53.866

On 9 Window, Poly 2 and RBF turn out to be most accurate amongst 5 Kernel functions, with the accuracies of 100.00 percent. Rest of functions showed accurate results in this order: Sigmoid, Normal and Poly 2. Unlike previous result, on 13 window, Poly 1 and RBF showed the highest accuracies, followed by Poly 2, Sigmoid, and Normal. On 17 window, both Poly 2 and RBF showed the most accurate results, followed by Poly 1, Sigmoid and Normal. Recognising the exceptional high accuracies of Poly 2 and RBF, I was able to confirm that CHIKV and DENV are different.

4.2. Decision tree

Several rules with frequencies above 0.83 were found on class 1 (DENV), and class 2 (CHIKV). According to Table 4, it can be assumed that position 11 is an important factor that differentiates the viruses since position 11 is the most frequent one. Simply putting, decision tree experiment results can indicate the possibility that CHIIV is derived from DENV.

Table 4: Decision Tree Rule Extraction with Frequencies

Class	**Rule on 9 window with Frequencies**	**Rule on 13 Widow with Frequencies**	**Rule on 17 Window with Frequencies**
DENV 9 Window	pos 4=T pos 7=K (0.833) pos 1= P pos 7=S (0.833) pos 3 = W (0.833) pos 5 = W (0.875)	pos 4 =W (0.889) pos 13=W (0.923) pos 12=W (0.875) pos 2 =W (0.857) pos 7=T pos 13=P (0.833)	pos 11=F (0.923) pos 11= L pos 17=K (0.875) pos 3= L pos 11 =E (0.833) pos 8= R (0.833) pos 11=M (0.900)
CHIKV 9 Window	pos 7=P pos 9=L (0.833) pos 1=C pos 7=L (0.857) pos 2= A (0.833)	pos 5=A (0.852) pos 10=Y pos 13=C (0.833)	pos 11=R pos 15=R (0.833) pos 11=S pos 15=T (0.833) pos 11=S pos 15=R (0.833) pos 1= P pos 8= R (0.833)

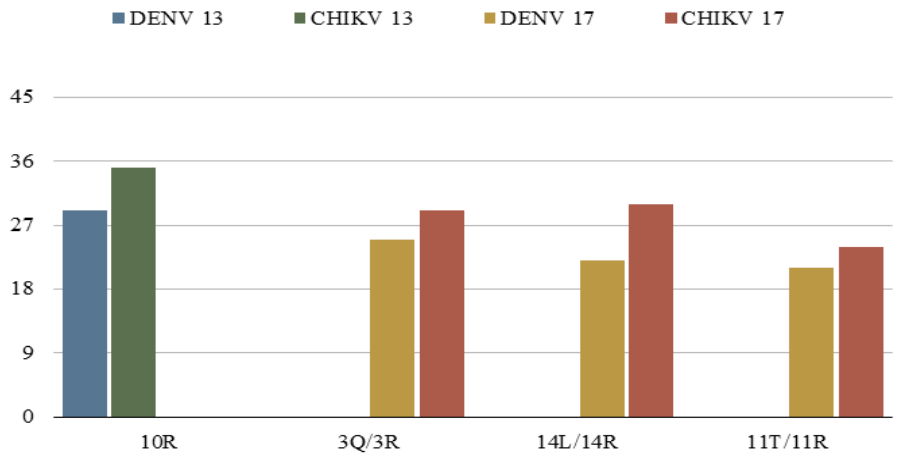

Fig. 5: Apriori Algorithm.

4.3. Apriori algorithm

Based on apriori algorithm results, I was able to compare numbers of amino acids embedded on the same position of DENV and CHIKV. Since there was no rule found on Dengue 9 window, I wasn't able to compare Dengue virus and Chikungunya virus on 9 Window. On 13 window, I compared numbers of 10R between Dengue virus and Chikungunya virus. On 17 window, I compared numbers of Dengue 3Q and Chikungunya 3R, Dengue 14L and Chikungunya 14R, and Dengue 11T and 11R. Results are shown on Fig 5.

5. Conclusion

The decision tree data showed several rules. From 9 window of Decision Tree, 46% of class 1 samples, (Dengue virus) are classified as class 2 (Chikungunya virus), while 45% of class 2 are classified as class 1. From 13window of Decision Tree, 36% of class 1 are classified to class 2, whilst 41% are classified to class 1. From 17window of Decision Tree, 35% of class 1 are classified to class 2, and 41% of class 2 are classified to class 1. Collectively, these results suggest that CHIKV and DENV share a common root. According to results from Apriori algorithm, there was comparatively large number of amino acid R, which is also known as an Arginine. Viruses containing an Arginine displays enhanced infectivity in mammalian cells but reduces infectivity in mosquito cells and dimities virulence in a mouse model of CHIKV disease [9]. This may indicate why clinical symptoms of Chikungunya include debilitating bilateral polyarthralgia and, in some cases, arthritis; whilst Dengue does not. On the other hand, amino acid T (Threonine) and amino acid S (Serine) were easy to find on Dengue virus. On Support vector machine experiments, Poly 2 classification showed exceptionally high accuracies on 9 window and 17 window. This indicates that CHIKV and DENV are different. To summarise, even though CHIKV and DENV can be classified, they still share common properties. The data I extracted from decision tree, apriori algorithm, and support vector machine require further research, yet I was able to find certain similarities and differences between CHIKV and DENV. I believe this research could help further research on genetic and clinical differentiation of CHIKV and DENV as well as accelerating the development of vaccines and medications.

6. References

[1] "Chikungunya." World Health Organization. N.p., n.d.

[2] "Dengue and Severe Dengue." World Health Organization. N.p., n.d.

[3] "Dengue." World Health Organization. N.p., n.d.

[4] Schmaljohn, Alan L. Alphaviruses (Togaviridae) and Flaviviruses (Flaviviridae). U.S. National Library of Medicine, n.d.

[5] Li, Xiao-Feng, Tao Jiang, Yong-Qiang Deng, Hui Zhao, Xue-Dong Yu, Qing Ye, Hong-Jiang Wang, Shun-Ya Zhu, Fu-Chun Zhang, E-De Qin, and Cheng-Feng Qin. "Complete Genome Sequence of a Chikungunya Virus Isolated in Guangdong, China." Journal of Virology. American Society for Microbiology, n.d.

[6] Henchal, E. A., and J. R. Putnak. "The Dengue Viruses." Clinical Microbiology Reviews. U.S. National Library of Medicine, n.d.

[7] Wasilewska, Professor Anita, and Lecture Note. APRIORI Algorithm (n.d.): n. pag. New York University. Anita Wasilewska.

[8] "Review: A Gentle Introduction to Support Vector Machines in Biomedicine." The Quarterly Review of Biology 88.4 (2013): 364. Web.

[9] Ashbrook, Alison W., Kristina S. Burrack, Laurie A. Silva, Stephanie A. Montgomery, Mark T. Heise, Thomas E. Morrison, and Terence S. Dermody. "Residue 82 of the Chikungunya Virus E2 Attachment Protein Modulates Viral Dissemination and Arthritis in Mice." Journal of Virology. American Society for Microbiology, n.d.