FINDING DIFFERENCE BETWEEN WEST AND EAST YELLOW FEVER BY APRIORI AND DECISION TREE

Jiwoo Jeon[1], Youngshin Joo, and Taeseon Yoon

[1]Department of National Science, Hankuk Academy of Foreign Studies, Yong-In, Republic of Korea

ABSTRACT

Yellow Fever is a fatal disease that causes yellowness and high fever. It is known to be highly contagious and it is mostly prevalent in Africa, where it is know n to be originated from. However, the derivations of this disease have reached other continents across the sea and they seem to be in different order. We mostly focused on the derivations within Africa and discovered that the geographical derivations of this disease also have different frequencies. We proved this phenomenon by analysing the sequences through apriori and decision tree.

KEYWORDS

Yellow fever virus, Apriori, Decision Tree, Transmission Cycle, Derivations

1. INTRODUCTION

Coquillettidia fuscopennata, more widely known as “yellow fever virus”(YFV) first broke out on the island of Barbados. Although vaccination is available now, it was considered to be lethal among the Europeans, opposed to the natives who were readily immune to the disease. The disease is currently most prevalent in tropical-like areas, with South America and Africa being the most vulnerable to its occurrence. Similarly, Ebola virus is a deadly disease threatening the nation. It first broke out in South Sudan, and ever since, several phylogeographic forms of Ebola virus have appeared, the genomic sequences differing slightly. Its vaccine has not yet been qualified, and studies are being continuously carried out. The most significant similarity between Ebola virus and the Yellow fever virus is that both diseases cause hemorrhagic fever. Regarding this feature, in this study, we investigated the genomic sequences of the proteins capable of such mechanism in both diseases.

2. RESULT METHOD

2.1. Yellow fever virus

Yellow fever is divided into 7 genotypes in including 2 West African genotype, 2 East Africa genotype, Angolan genotype and 2 South American genotype.[1,2,3-5]. Specifically West Africa genotype I (Nigeria, Cameroon, and Gabon), West Africa genotype II (Senegal, Guinea, Ivory Coast, and Ghana), East and Central African genotype (Sudan, Ethiopia, Central African Republic, and Democratic Republic of Congo), East African genotype (Kenya), and Angola genotype (Angola).

In our research, we will find the difference and expand relationship among West African(West Africa genotype I (Ivory Coast), West Africa genotype II (Senegal Asibi, FVV, Gambia01)), east African(Uganda48a), east/central African(Ethiopia61b) and Angolan (Angola71) genotypes.

Interestingly in West Africa, outbreaks happened more often, while East Africa outbreak is uncommon. The cause for this can be found in the genetic variability of the virus. the West African genotype I whic shows a higher heterogeneity than the West African genotype II or the East/Central African genotype. (4) Also, the cause can be found in host adaptation, host behavior, climatic and ecological factors.(6)

Strain	Origin	Date collected	Source	Genotype	Reference (GenBank accession no.)
Asibi	Ghana	1927	Human	West Africa II	Hahn et al. (1987) (AY640589)
14FA	Angola	1971	Human	Angola	Mutebi et al. (2001) (AY968064)
FVV	Senegal	1927	Human	West Africa II	Wang et al. (1995) (U21056)
Gambia01	Gambia	2001	Human	West Africa II	Colebunders et al. (2002) (AY572535)
85-82H	Ivory Coast	1982	Human	West Africa I	Pisano et al. (1997) (U54798)
A7094A2	Uganda	1948	Unknown	East Africa	Mutebi et al. (2001) (AY968065)
Couma	Ethiopia	1961	Human	East/central Africa	Serie et al. (1968) (DQ235229)

Table 1. Yellow fever viruses [5]

In the present study, we used the complete yellow fever virus genome from NCBI genback. The table above shows the viruses we used. We used one West Africa I, East Africa, East/central Africa, Angola virus and 3 West Africa II viruses.

The strain which has West Africa I genotype is 85-82H which is from Ivory coast in 1982. East Africa's strain is named A7094A2 which is from Uganda in 1948, East/central Africa's strain is called Couma which is from Ethiopia in 1961, and Angola is called 14FA which is from Angola in 1971. The virus which has the genotype of West Africa II is Asibi, FVV, and Gambia 01 which is each from Ghana, Senegal and Gambia and collected in 1927, 1927 and 2001.

2.1. Apriori Algorithm

In our research, we used apriori algorithm to find the relation between West, East and Angolan genotype of yellow fever virus. Apriori algorithm finds out the association rule between the data based on the frequency.

This algorithm counts the frequency of the item to determine the large itemsets. [7] Candidate generation works, which extends frequent subsets one at a time then each candidate is tested against the extension until no further data is found. [8] The form of the algorithm is similar as figure 1.

1) L_1 = {large 1-itemsets};
2) **for** ($k = 2$; $L_{k-1} \neq \emptyset$; k++) **do begin**
3) C_k = apriori-gen(L_{k-1}); // New candidates
4) **forall** transactions $t \in \mathcal{D}$ **do begin**
5) C_t = subset(C_k, t); // Candidates contained in t
6) **forall** candidates $c \in C_t$ **do**
7) c.count++;
8) **end**
9) $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
10) **end**
11) Answer = $\bigcup_k L_k$;

Figure 1. Apriori algorithm

We divided the amino acid sequences of virus into 9, 13, 17 windows and analyzed the frequent sequence pattern respectively.

2.2 Decision Tree

A decision tree classifies data items (Fig. 1a) by posing a series of questions about the features associated with the items. Each question is contained in a node, and every internal node points to one child node for each possible answer to its question. The questions thereby form a hierarchy, encoded as a tree. In the simplest form (Fig. 1b), we ask yes-or-no questions, and each internal node has a 'yes' child and a 'no' child. An item is sorted into a class by following the path from the topmost node, the root, to a node without children, a leaf, according to the answers that apply to the item under consideration. An item is assigned to the class that has been associated with the leaf it reaches. In some variations, each leaf contains a probability distribution over the classes that estimates the conditional probability that an item reaching the leaf belongs to a given class.

	Leaves		Width		Height	
	J48	VTJ48	J48	VTJ48	J48	VTJ48
anneal	37.69	12.98	2753.62	2555.43	670.11	677.68
anneal.ORIG	46.37	11.10	3426.41	1362.30	868.05	546.30
arrhythmia	40.59	10.20	1679.34	1589.04	1555.57	1462.47
audiology	30.25	9.11	3799.18	3781.91	923.98	921.00
autos	45.25	12.77	6527.37	4199.37	654.02	637.58
balance-scale	41.24	25.86	1986.12	1222.98	821.96	747.91
breast-cancer	9.60	4.04	1177.92	1518.04	348.63	354.23
breast-w	12.08	14.23	781.99	967.01	637.84	698.75
colic	6.07	8.76	546.44	1198.39	360.41	424.61
colic.ORIG	1.00	6.83	1.00	480.83	1.00	372.96
credit-a	21.40	12.01	1664.81	1098.50	669.91	619.21
credit-g	89.05	7.07	13906.86	1077.07	877.89	335.60
diabetes	21.87	11.97	1488.65	963.30	830.31	694.87
ecoli	18.70	17.78	1039.47	1055.72	735.37	723.22
glass	23.73	12.23	2293.10	1838.96	827.88	754.41
heart-c	36.05	8.85	3273.48	1300.71	618.10	476.47
heart-h	7.21	8.17	673.28	1042.54	408.37	464.92
heart-statlog	17.85	13.41	1577.55	1309.66	633.84	605.13
hepatitis	9.24	12.41	522.66	754.78	571.69	659.98
hypothyroid	14.39	13.43	1101.64	1070.54	756.02	771.15
ionosphere	13.85	11.59	1070.98	1019.02	775.47	734.02
iris	4.69	4.76	227.87	231.10	428.43	432.21
kr-vs-kp	28.98	13.16	1187.64	1104.25	1091.28	1076.77
labor	4.00	5.20	329.00	464.17	333.06	380.56
letter	1165.00	12.65	63285.55	63344.28	1916.69	1919.54
lymph	17.43	10.12	1863.97	1252.03	580.12	462.35
mushroom	24.93	24.93	1022.25	1022.25	527.00	527.00
optdigits	205.46	16.09	11154.56	11195.65	1330.36	1334.04
pendigits	166.13	16.94	10719.41	10764.98	1297.09	1296.05
primary-tumor	43.18	14.62	3794.33	1797.86	891.43	789.16
segment	41.12	11.09	3749.02	3748.84	1084.95	1085.78
sick	27.59	14.22	1763.57	1087.54	815.68	710.67
sonar	14.71	13.80	1107.13	1089.68	665.59	659.63
soybean	61.28	11.04	6175.62	6180.02	913.67	920.67
splice	173.83	20.78	7537.58	6176.44	759.48	731.51
vehicle	69.22	16.27	5069.70	4183.99	1168.31	1065.60
vote	5.81	6.22	390.94	432.98	508.98	513.86
vowel	126.41	10.58	11046.43	11045.28	985.60	986.01
waveform-5000	295.66	16.82	16325.97	13756.92	1494.51	1386.66
zoo	8.31	8.31	436.69	436.69	567.50	567.50

doi:10.1371/journal.pone.0033812.t003

Figure 2 a simple decision tree model

3. RESULTS

2.1. Apriori Algorithm

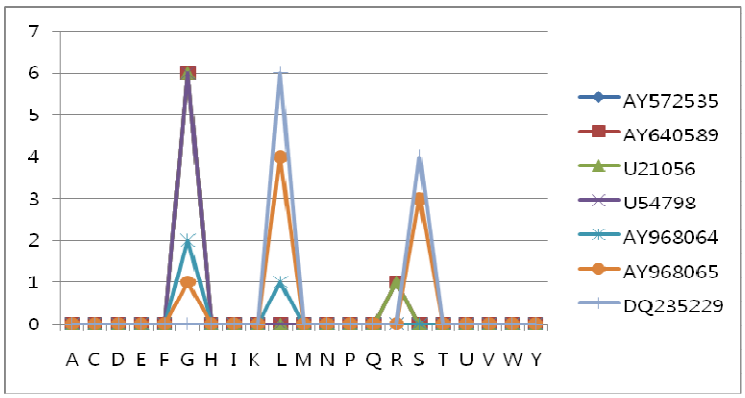

Figure 3 Apriori algorithm window 9

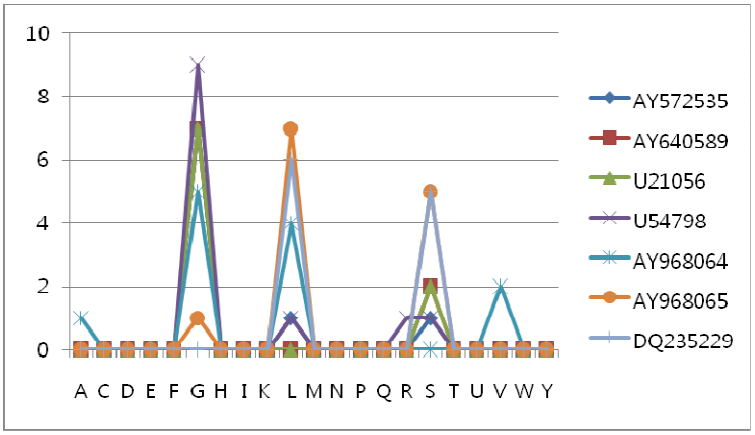

Figure 4 Apriori algorithm window 13

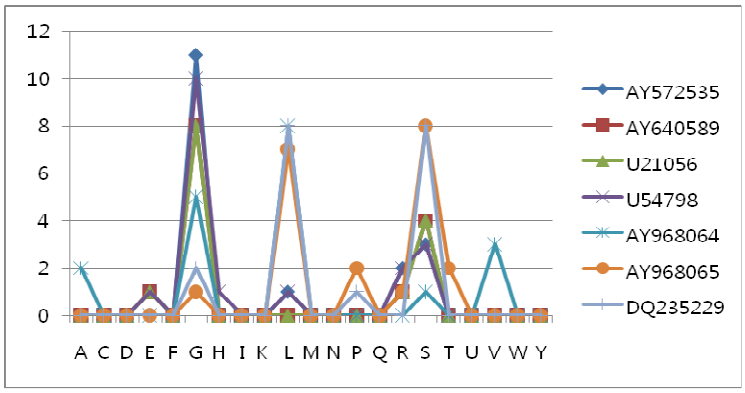

Figure 5 Apriori algorithm window 17

The graph above shows the frequency of the amino acids of yellow fever virus. From the left the genotype of Gambia, Asibi, Senegal is West Africa II, Ivory coast is West Africa I, Angola's genotype is Angola and Uganda and Ethiopia's genotype is East Africa. To Compare those two viruses.

In all 9, 13, 17 window we could find out that West Africa I, II had no big difference. However, West Africa and East Africa had big difference. West Africa showed glycine the most and rarely showed arginine and serin. Whereas, East Africa showed serin and leucine the most and rarely showed proline and glycine. Also angola virus showed leucine, glycine the most and rarely showed alanine and valine.

Therefore as the reference had shown that in west Africa outbreaks happen more often cause of the genetic variability and higher heterogeneity of the virus, we are predicting that amino acids above could help the outbreak of the virus.

2.2. Decision Tree

Window	Rule	Frequency
Window 9	pos4 = E pos7 = E pos3 = T pos7 = A pos8 = V pos7 = F pos9 = K pos7 = A pos8 = M pos1 = D pos6 = S pos1= Y pos5 = T pos7 = M pos9 = E pos4 = E pos7 = E pos3=M pos7 = E pos7 = F pos9 = K pos3 = V pos4 = I pos7= M pos9 = E pos4 = E pos7 = E pos7 = M pos9 = E pos4 = E pos7 = E pos4 = E pos7 = E pos6 = V pos7 = A pos4 = E pos7 = E pos7 = K pos7 = M pos9 = E	0.800
	pos6 = M pos7 = L pos6 = M pos7 = L pos1= A pos7 = G pos1= A pos7=G pos1 = A pos7 = G pos1 = M pos8 = V	0.833
	pos4 = A pos7 = E	0.857
Window 13	pos6 = M pos7 = L pos6 = M pos7 = L pos1= A pos7 = G pos1= A pos7=G pos1 = A pos7 = G pos1 = M pos8 = V pos8 = A pos11= E pos7 = L pos12 = K pos8 = G pos11 = Y pos7 = L pos12 = K	0.833
	pos4 = A pos7 = E pos8 = L pos11 = M pos1 = G pos6 = E	0.857

	pos11 = V pos12 = I pos3 = K pos12 = R pos1 = E pos12 = T pos11 = V pos12 = I pos3 = M pos11 = L pos3 = T pos7 = L pos12 = K pos3 = K pos12 = R pos12 = V pos1 = V pos11 = S pos3 = E pos13 = E pos7 = L pos12 = K pos3 = A pos11 = A pos3 = M pos11 = L pos3 = T pos12 = V pos4 = D pos11 = S pos8 = Y pos13 = V pos8 = P pos12 = P p os3 = M pos11 = L pos3 = T pos12 = V pos1 = V pos11 = S pos3 = E pos13 = E pos1 = A po s12 = T pos1 = E pos12 = T	0.800
Window17	pos2 = A pos8 = V pos2 = Y pos10 = L pos2 = Y pos10 = L	0.833
	pos2 = K pos2 = L pos2 = R pos11 = M pos2 = T pos4 = A pos2 = T pos17 = V pos14 = M pos2 = R pos11 = M pos17 = V pos1 = V pos2 = R pos2 = N pos1 = L pos8 = T pos4 = M pos2 = R pos2 = L pos14 = A pos4 = E pos4 = A pos11 = M pos2 = E pos4 = E pos7 = M pos14 = A pos2 = L pos4 = A pos4 = E pos14 = A pos2 = R pos8 = T pos1 = V pos2 = N	0.800

Figure 6 Decision tree of angola virus

Window	Rule	Frequency
Window 17	pos3 = S pos4 = T pos2 = K pos13 = R pos2 = K pos13 = R pos3 = S pos4 = T pos2 = K pos2 = K pos13 = R pos13 = R pos3 = S pos4 = T pos2 = K pos13 = R	0.833
	pos8 = Rpos11 = Rpos3 = Ipos3 = S pos7 = Mpos4 = T	0.800

Figure 7 Decision Tree of uganda virus

Window	Rule	Frequency
Window 13	pos7 = Q pos12 = S	0.800

Figure 8 Decision Tree of Ethiopia Virus

Window	Rule	Frequency
Window 13	pos6 = L pos12 = G pos13 = S	0.800

Figure 9 Decision Tree of Ivory Coast virus

Upon analyzing the statistics of this decision tree we used on Yellow fever, we assumed that there must be some reasons behind the cluster of information around class 3. So we continued to analyze the rules one by one and decided there must be some atypical properties to this disease that can explain such phenomenon.

4. CONCLUSIONS

Our study reveals a dominant distinction between West Africa yellow fever and East/Central Africa regarding the analyses of the genome. As shown in the apparent disparity between the frequency of different amino acids, we inferred that this might entail to different properties between West and East/Central yellow fever. According to various studies, all 7 genotypes of Yellow Fever (5 of them in Africa) display the same symptoms; facial tone turning yellow and acute liver phase. However there also seem to be roughly 3 characteristics that distinguish the geographic genotypes of yellow fever.

1. Studies to date show that Yellow Fever first evolved from East/ Central Africa and branched out to West Africa and was later transported to South America. From analyses of the genome, West Africa virus strain displayed high similarity with the South American strain, but comparatively lower similarity with the East and central African strain despite the fact that they derived from the same continent.

2. Three different transmission cycles exist in Africa ; jungle(sylvatic), intermediate(savannah) and urban. As the name follows, jungle cycle occurs in tropical rain forests, intermediate cycle in savannah, and urban cycle via travellers introducing the disease to unprecedented areas. The vectors differ,too, as it can be shown in the following diagram.

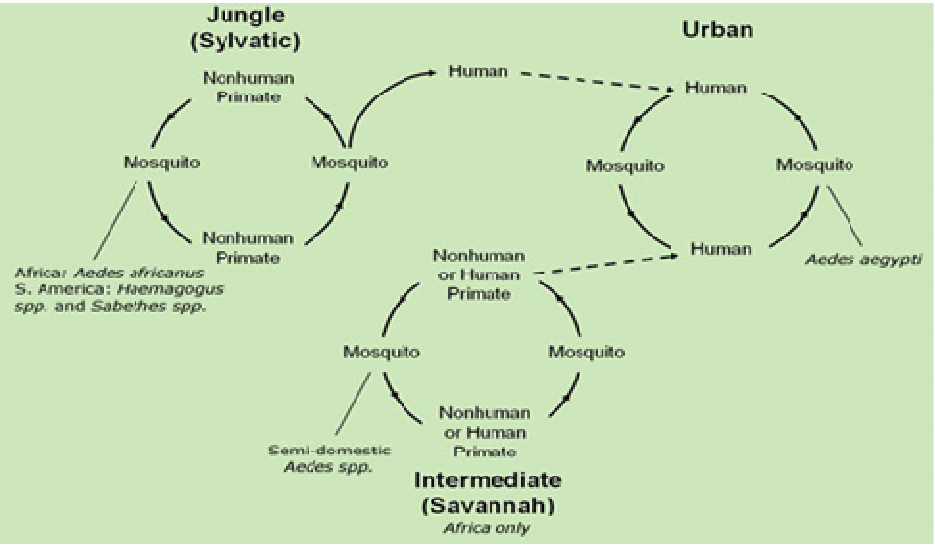

Figure 10. Transmission Cycles

3. West African strains have a 3'NCR of 511 nucleotides while strains from Central/East Africa had shorter 3'NCRs (443-469 nucleotides) due to the absence of YF specific repeat sequences (RYFs). Having such qualitative difference within a single disease seems to advocate the idea that the results shown in our study are meaningful.

REFERENCES

[1] Mutebi JP, Barrett AD. 2002. The epidemiology of yellow fever in Africa. Microbes Infect. 4:1459 – 1468.

[2] Barrett AD, Monath TP. 2003. Epidemiology and ecology of yellow fever virus. Adv. Virus Res. 61:291–315

[3] de Souza RP, Foster PG, Sallum MA, Coimbra TL, Maeda AY, Silveira VR, Moreno ES, da Silva FG, Rocco IM, Ferreira IB, Suzuki A, Oshiro FM, Petrella SM, Pereira LE, Katz G, Tengan CH, Siciliano MM, Dos Santos CL. 2010. Detection of a new yellow fever virus lineage within the South American genotype I in Brazil. J. Med. Virol. 82:175–185.

[4] Mutebi JP, Wang H, Li L, Bryant JE, Barrett AD. 2001. Phylogenetic and evolutionary relationships among yellow fever virus isolates in Africa. J. Virol. 75:6999 –7008.

[5] von Lindern JJ, Aroner S, Barrett ND, Wicker JA, Davis CT, Barrett AD. 2006. Genome analysis and phylogenetic relationships between east, central and west African isolates of Yellow fever virus. J. Gen. Virol. 87: 895–907.

[6] Nina K. Stock,a Hewad Laraway,a Ousmane Faye,b Mawlouth Diallo,b Matthias Niedrig,a Amadou A. Sallb, 2013, Biological and Phylogenetic Characteristics of Yellow Fever Virus Lineages from West Africa, 87 Journal of Virology p. 2895–2907

[7] R. Agrawal and R. Srikant, "Fast Algorithms for mining association rules", VLDB, pp.487-499, Sept, 1994.

[8] Eunby Go, Seungmin Lee, and Taeseon Yoon, Analysis of Ebolavirus with Decision Tree and Apriori algorithm, International Journal of Machine Learning and Computing, Vol. 4, No. 6, December 2014

Authors

Ji woo Jeon

Was born in Korea, in 1997. She is a junior, eleventh grade student in the national science program at Hankuk Academy of Foreign studies since 2014.

Youngshin Joo

Was born in Korea, in 1997. She is a junior, eleventh grade student in the national science program at Hankuk Academy of Foreign studies since 2014.

Taeseon Yoon

was born in Seoul, Korea, in 1972. He was Ph.D. Candidate 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. Since December 2004, he has been with the Hankuk Academy of Foreign Studies, where he was a Computer Science and Statistics Teacher.