COMPARISON OF NATIVE AND INVASIVE SPECIES OF AEDES ALBOPICTUS USING DATA MINING; DERIVING THEIR GENETIC VARIATION FROM COI SEQUENCE VARIABILITY

Beomsu Kim [1] and Taeseon Yoon[2]

[1,2]Department of International, Hankuk Academy of Foreign Studies,
Yong-In, Republic of KOREA

ABSTRACT

In this study, Apriori algorithm was employed to genetic sequence patterns of Aedes albopictus (Asian tiger mosquito, Skuse, Dengue fever vector) in native and invasive ranges. It has spread widely across the world from its native region, South-east Asian countries, to every continent except Antarctica. To find the congruity of the species and their origins, I decided to compare them. To analyze the short-term genetic mutations, mtDNA was used to derive the pattern.

KEYWORDS

Aedes albopictus, Cytochrome Oxidase I, mtDNA, Biological invasion, Bioinformatics, Apriori Algorithm

1. INTRODUCTION

Known as the Asian tiger mosquito, or Skuse, Ae. Albopictus is currently the most prevalent invasive mosquito species compared to other mosquito species. It has gone under dramatic global expansion facilitated by human activities, in particular the movement of used tires and „lucky bamboo" [1]. The global distribution of Ae. Albopictus has resulted from being transported along with public and private transport. It ranks in the top 100 species in the invasive species list of the Invasive Species Specialist list [3]. Its first observation was in Albania, 1979. It has spread northward and eastward from its first occurrence Texas, USA in 1985. It is now reported to be in over 25 US states. Its first occurrence in Latin America was in Brazil, 1986. Another report was filed from Mexico in 1988 [7]. Due to various attributes of Ae. Albopictus, including environmental adaptability, high survival rate, lack of check, and no efficient control system, it was able to proliferate in foreign regions. For the past decades, dengue infection has increased dramatically. Current heath reports are misclassified and may have underestimated the numbers. One estimate indicates 390 million dengue infection per year [2]. Moreover, because of its aptitude of feeding on various hosts, it has a potential to serve as a vector of transferring zoonotic pathogens to human population [4]. The mtDNA was used in comparing the different populations, because unlike nuclear DNA, it undergoes the discrete process of relaxed replication.

This indicates that point mutations or indels cumulate through the lifetime of a multicellular individual [5]. Data mining technique can be used to compare the mtDNA sequences since the mutations in different populations accumulate over time, providing a larger dataset. Association rules can be applied to reveal associations between gene sequences of different species or populations [6]. Using this inference, I implemented the apriori algorithm which exploits the fact that no superset of an itemset having not enough support can have enough support [8]. By comparing the COI Aerospace Engineering: An International Journal (AEROIJ), Vol. 1 , No.1 , 2015 10 sequences of different populations with Apriori algorithm and Shannon entropy, these samples could be analysed for similarities and differences.

2. METHOD

2.1. Using an apriori algorithm to acquire the number of several types of amino acids.

I used the apriori algorithm to analyze the similarity between the native aedes albopictus and invasive ae. albopictus. When using the apriori algorithm, I used two options; 5- window, and 7-window. 5-window means we separated the genomic sequence into 5 parts and found out the associations and similarities between different genes (base sequences). Predictive Apriori, which analysed large gene data sets via Apriori algorithm, was used for the experiments. This algorithm uses two concepts, support (the percentage of the population that satisfies the rule), and confidence (the percentage that the consequence is also satisfied when the antecedent exists). These two concepts are used to derive rules from the dataset. Fig. 1 summarizes the Apriori algorithm explained above Figure 1. Apriori Algorithm There is a certain tendency that as the number of windows in each experiment increases, the number of rules generated differs more and more with that of an experiment with lower number of windows. This can be explained by the characteristic of the Apriori algorithm. As the number of window increases, the number of instance for the support required increases accordingly. The base sequence sequences were acquired from www.ncbi.nlm.nih.gov To gain a wider view on the genetic variance in relation to the proximity to the native region, the native samples were selected from Malaysia and East Timor while the invasive samples were from Australia and California.

Ae. albopictus voucher ALBnzmk36 COI gene
http://www.ncbi.nlm.nih.gov/nuccore/499384591
Ae. albopictus voucher ALBnzmk37 COI gene
https://www.ncbi.nlm.nih.gov/nuccore/499384593
Ae. albopictus haplotype ET197 COI gene
https://www.ncbi.nlm.nih.gov/nuccore/523541290
Ae. albopictus haplotype ET199 gene
http://www.ncbi.nlm.nih.gov/nuccore/523541294
Ae. albopictus haplotype Eru50 COI gene
https://www.ncbi.nlm.nih.gov/nuccore/523540710
Ae. albopictus haplotype Eru28 COI gene
http://www.ncbi.nlm.nih.gov/nuccore/523540708
Ae. albopictus haplotype H29 gene
http://www.ncbi.nlm.nih.gov/nuccore/519899545
Ae. albopictus haplotype H30 COI gene
https://www.ncbi.nlm.nih.gov/nuccore/519899547

2.2 Using the Shannon Entropy theorem to calculate uncertainty probability associated with random variables.

Shannon entropy presents one of the most valuable insights into the information theory. Entropy measures the uncertainty associated with a random variable, such as the value of information in a nucleotide sequence. The concept was introduced by Claude E. Shannon in the paper "A Mathematical Theory of Communication" (1948). Shannon entropy allows the estimation of average minimum number of bits needed to encode a string of symbols based on the alphabet size and frequency of the symbols.

2.3. Using the decision tree to derive differences between each mtDNA gene sequences.

Decision tree learning is an analysis method commonly used to derive the model that predicts the value of a target variable based on several input variables [9]. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree stands for some possible attributes of the input data and each branch corresponds to a possible class for the instance [10].

C5.0, a decision tree program, was used to train and run the algorithm. After running the experiment, cross validation is used to get a more reliable estimate of predictive accuracy. In this experiment, 10-fold validation using ruleset classifiers was made. This was also conducted in two type: 5-window and 7-window.

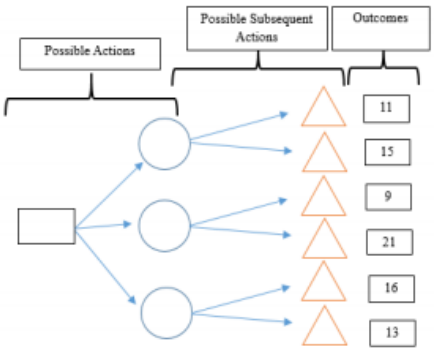

Figure 2. A Basic Decision Tree

3. EXPERIMENT

By using the mtDNA dataset and extracting meaningful areas which encode amino acids and contribute in protein translation, I found significant similarities. In each analysis, the results took the following form:

1. amino4 => L 7
2. amino5=> V 6

3. amino3=> S 3

The results are the rules attainted from Ae. albopictus voucher ALBnzmk37 COI gene, extracted in Malaysia, a native region, using 5-window. “Amino4 = L” means in the 7-window, the “L” amino acid takes the fourth space, and will be translated the fourth among 7 other amino acids in 12 the same window. The number “7” means there are 7 amino2=L that has passed the threshold that was initially indicated by the algorithm. After getting a result from the apriori algorithm, we counted the different kind of amino acids and made a table of them. The position of the amino acids were put into consideration. First, these are tables that show the number of amino acids from the 5-window experiment. Sample from East Timor generated 27 rules, sample from Malaysia 21 rules, Australia 27, and California 16. The number of rules derived differs for itemsets that satisfy the given threshold, or the support and confidence, varies with each dataset.

Table 1. THE NUMBER OF AMINO ACIDS IN AE. ALBOPICTUS FROM 5WINDOW EXPERIMENT.

Sample Region	{Position, Amino Acid (frequency)}
East Timor	{2, L (7)}, {4, L (7)}, {5, V (6)}, {1, L (5)}, {2, G (5)}, {2, I (5)}, {3, I (5)}, {1, I (4)}, {2, P (4)}, {3, T (4)}, {4, I (4)}, {4, T (4)}, {5, I (4)}, {5, S (4)}, {1, A (3)}, {1, F (3)}, {1, G (3)}, {1, S (3)}, {2, N (3)}, {3, A (3)} {3, L (3)}, {3, S (3)}, {4, A (3)}, {4, P (3)}, {4, S (3)}, {5, A (3)}, {5, L(3)}
Malaysia	{1, G (6)}, {1, L (6)}, {4, V (6)}, {5, L (6)}, {1, I (5)}, {2, T (5)}, {3, L (5)}, {3, T (5)}, {4, S (5)}, {2, I (4)}, {2, L (4)}, {4, I (4)}, {4, L (4)}, {5, G (4)}, {2, A (3)}, {3, A (3)}, {3, I (3)}, {3, P (3)}, {3, S (3)}, {5, F (3)}, {5, S (3)}
Australia	{2, L (7)}, {4, L (7)}, {5, V (6)},{1, L (5)}, {2, G (5)}, {2, I (5)}, {3, I (5)}, {1, I (4)}, {2, P (4)}, {3, T (4)}, {4, I (4)}, {4, T (4)}, {5, I (4)}, {5, S (4)}, {1, A (3)}, {1, F (3)}, {1, G (3)}, {1, S (3)}, {2, N (3)},{3, A (3)}, {3, L (3)}, {3, S (3)}, {4, A (3)}, {4, P (3)}, {4, S (3)}, {5, A (3)}, {5, L (3)}
California	{2, I (17)}, {4, I (17)}, {5, I (16)}, {3, I (15), }, {1, I {14}}, {4, L (14)}, {5, L (14)}, {5, F (13)}, {2, G (12)}, {1, T (11)}, {3, T (11)}, {1, G (10)},{1, L (10)}, {2, L (10)}, {3, L (10)}, {4, G (10)}

From the 7-Window, sample from East Timor generated 22 rules, Malaysia 33 rules, Australia 22 rules, and California 21 rules.

Table 2. THE NUMBER OF AMINO ACIDS IN AE. ALBOPICTUS FROM 7WINDOW EXPERIMENT

Sample Region	{Position, Amino Acid (frequency)}
East Timor	{1, L (6)}, {3, I (6)}, {7, L (6)}, {4, S (5)}, {1, A (4)}, {2, G (4)}, {2, L (4)}, {3, L (4)}, {5, L (4)}, {5, S (4)}, {6, I (4)}, {1, G (3)}, {1, I (3)}, {1, V (3)}, {2, V (3)}, {4, G (3)}, {5, I (3)}, {6, P (3)}, {7, I (3)}, {7, T (3)}, {2, A (2)}, {2, P (2)}
Malaysia	{5, L (7)}, {4, L (6)}, {1, S (5)}, {2, L (4)}, {5, A (4)}, {6, G (4)}, {7, I (4)}, {7, L (4)}, {1, G (3)}, {2, I (3)}, {2, S (3)}, {3, D (3)}, {3, I (3)}, {3, P (3)}, {4, I (3)}, {4, T (3)}, {6, I (3)}, {6, L (3)}, {7, G (3)}, {7, T (3)}, {1, F (2)}, {1, H (3)}, {1, N (2)}, {1, T (2)}, {1, A (2)}, {1, F (2)}, {2, V (2)}, {3, F (2)}, {4, P (2)},{4, S (2)}, {4, V (2)}, {5, G (2)}, {5, N (2)}
Australia	{1, L (6)}, {3, I (6)}, {7, L (6)}, {4, S (5)},{1, A (4)}, {2, G (4)}, {2, L (4)}, {3, L (4)},{5, L (4)},{3, S (4)}, {6, I (4)}, {1, G (3)}, {1, I (3)}, {1, V (3)}, {2, V (3)}, {{4, G (3)},{5, I (3)}, {6, P (3)},
	{7, I (3)}, {7, T (3)}, {2, A (2)}, {2, P (2)}
California	{3, I (14)}, {2, L (13)}, {7, I (13)}, {4, I (12)}, {1,G (11)}, {2, I (11)}, {6, G (11)}, {1, I (10)}, {4, L (10)}, {5, I (10)}, {5, L (10)}, {6, I (9)}, {7, F (9)}, {7, L (9)}, {5, F (8)}, {6, L (8)}, {1, S (7)}, {3, T (7)}, {4, F (7)}, {5, A (7)}, {7, T (7)}

After counting the amino acids and their association rules, Shannon entropy was used to compare the similarities of each data. For b, 2 (bit) was used. The uncertainty probability associated with random variables was calculated: Table

$$H = -\sum_{i} p_i \log_b p_i$$

3. THE SHANNON ENTROPY H(X) VALUES FROM 5WINDOW EXPERIMENT.

Sample Region	H(x)
Malaysia	4.11346
East Timor	4.69413
California	3.97235
Australia	4.69413

Table 4. THE SHANNON ENTROPY H(X) VALUES FROM 7WINDOW EXPERIMENT.

Sample Region	H(x)
Malaysia	4.88118
East Timor	4.39496
California	4.28937
Australia	4.39496

In the case of the decision tree, although over 30 rules were generated for each fold, for the sake of conciseness, I have displayed rules that had gained confidence over 0.7. Overall error indicates the percentages of misclassification. The Class in the index stands for the class that the amino acid can belong to. Class 1,2 is gene from Australia, 3,4 from California, 5, 6 from East Timor, and 7, 8 from Malaysia.

Table 5. SUMMARY OF THE 5WINDOW EXPERIMENT.

Fold Number	Rules {Position, Amino Acid, Class}	Overall Error (%)
1	{1, T, 2} {4, T, 2}, {5, L, 2} {1, E, 7} {5, N, 7} {1, D, 8} {5, N, 8}	84.9
2	{1, E, 7} {2, I, 7} {1, D, 8} {5, N, 8}	93.2
3	{1, T, 2} {4, T, 2} {5, L, 2} {1, E, 7} {5, N, 7} {1, I, 7} {4, S, 7} {1, D, 8} {5, N, 8}	90.4
4	{1, T, 2} {2, G, 2} {3, I, 2} {1, E, 7} {2, I, 7} {1, D, 8} {5, N, 8} {1, S, 8} {3, P, 8}	89.0
5	{1, E, 7} {2, I, 7} {1, S, 8} {3, P, 8} {1, D, 8} {5, N, 8} {2, L, 8} {4, S, 8}	91.8
6	{1, E, 7} {2, I, 7} {1, D, 8} {5, N, 8} {1, S, 8} {3, P, 8}	95.9
7	{1, S, 8} {3, P, 8} {1, D, 8} {5, N, 8}	91.8
8	{1, T, 2} {4, T, 2} {1, E, 7} {5, N, 7} {1, S, 8} {5, L, 8} {1, D, 8} {5, N, 8}	90.4
9	{1, E, 7} {2, I, 7} {1, I, 7} {2, T, 7} {3, P, 7} {1, D, 8} {5, N, 8}	97.3
10	{1, E, 7} {2, I, 7} {1, S, 8} {3, P, 8} {1, D, 8} {5, N, 8}	89.2

Table 6. SUMMARY OF THE 7WINDOW EXPERIMENT

Fold Number	Rules {Position, Amino Acid, Class}	Overall Error (%)
1	{1, T, 2} {2, G, 2} {4, T, 2}	90.6
2	{1, T, 2} {2, G, 2} {4, T, 2}	92.5
3	{1, T, 2} {2, G, 2} {4, T, 2} {1, E, 7} {4, L, 7}	94.3
4	{1, T, 2} {2, G, 2} {4, T, 2}	90.6
5	{1, T, 2} {2, G, 2} {4, T, 2}	90.6
6	{1, T, 2} {2, G, 2} {4, T, 2} {1, I, 4} {4, A, 4} {1, E, 7} {4, L, 7}	94.3
7	{1, T, 2} {2, G, 2} {4, T, 2} {1, G, 4} {4, L, 4}	98.1
8	{1, T, 2} {2, G, 2} {4, T, 2}	94.4
9	{1, T, 2} {2, G, 2} {4, T, 2}	90.7
10	{1, T, 2} {2, G, 2} {4, T, 2}	92.6

4. RESULTS AND DISCUSSION

Through this experiment, the features of Cytochrome Oxidase Subunit I mtDNA of ae. albopictus were figured out and analyzed for each type of samples: native species from Malaysia and East Timor, invasive species from Australia and California. First main experiment, in using experimental method 1, Apriori algorithm, division into 2 factions called window 5 and window 7 was made. Overall 8 experiments were performed to differentiate each population of ae. albopictus. The results from both 5Window and 7Window experiments showed clear characteristics. The sample from East Timor and showed clear similarity. The samples from East Timor and Australia, however, showed clear differences in COI structure from each other. For the second experiment, by calculating the uncertainty probability associated with random variables using Shannon entropy theorem, the amino acid sequence was contrasted. As results turned out, from 5-Window experiment and 7-Window experiment in that order, Shannon entropy for ae. albopictus from Malaysia is H(x)=4.27102 and H(x)=4.94556. East Timor is H(x)=4.69413 and H(x)=4.39496. California is H(x)=3.97235 and H(x)=4.35789. Finally, Australia is H(x)=4.69413 and H(x)=4.39496. The Shannon entropy of East Timor and Australia showed identical values with each other for data from both 5Window experiment and 7Window experiment. Looking at the increments and decrements, the sample from Malaysia showed a large variation of approximately 0.8, indicating a large change in the uncertainty associated with random variables in data from the 5Window and 7Window experiment. In training the decision tree, the results corroborated the conjectures that were made based on the results from the Apriori algorithm and the Shannon entropy. Although the 5window experiment showed some variances within groups, 7window showed less diversity due to stricter standards for datasets.Contrary to the expectation 15 that the samples from two native regions, Malaysia and East Timor would show more similarity than the other two from foreign regions, samples from East Timor and Australia showed identical results from both the Apriori experiment and Shannon entropy analysis, leading to the conclusion that the species from East Timor and Australia share similar or identical

Cytochrome Oxidase Subunit I gene sequences, either due to their geographical proximity or recentness of the invasion to Australia from East Timor. albopictus.

5. CONCLUSION

The steps taken towards stopping the invasion of aedes albopictus has been so far unsuccessful. Their remarkable adaptability to foreign environments, proximity to humans, and the reproductive biology has led to their proliferation not only in native regions but also in foreign regions such as Australia and California in this case. In looking for a way to containing ae. albopictus, Apriori algorithm and the Shannon entropy was used to compare each types. Additional experiments with decision tree was used to back up this result. Through analysis with various approaches, it was discovered that some samples from different regions have remarkably similar COI gene sequence, leading to the conjecture that identical methods could be used to control the population in both regions. This presents further insight that samples from more diverse regions could be compared and contrasted by the method used in this experiment, expediting the process of ae. albopictus containment.

REFERENCES

[1] Zhong D, Lo E, Hu R, Metzger ME, Cummings R, et al. (2013) Genetic Analysis of Invasive Aedes albopictus Populations in Los Angeles County, California and Its Potential Public Health Impact. PLoS ONE 8(7): e68586. doi:10.1371/journal.pone.0068586

[2] Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL et.al. The global distribution and burden of dengue. Nature;496:504-507.

[3] Buhagiar JA. A second record of Aedes (Stegomyia) albopictus (Diptera: Culicidae) in Malta. European Mosquito Bulletin. 2009;27:65-7.

[4] Benedict MQ, Levine RS, Hawley WA, Lounibos LP. Spread of the tiger: global risk of invasion by the mosquito Aedes albopictus. Vector Borne Zoonotic Dis. 2007 Spring;7(1):76-85.

[5] Filipe Pereira, João Carneiro, Barbara van Asch. A Guide for Mitochondria DNA Analysis in NonHuman Forensic Investigations. The Open Forensics Science Journal. 2010 Mar. 12p.

[6] Chad Creighton, Samir Hanash. Mining gene expression databases for association rules. Oxford University Press. 2003. 8p.

[7] Manorenjitha Malar A/P Sivanathan. THE ECOLOGY AND BIOLOGY OF Aedes aegypti (L.) AND Aedes albopictus (Skuse) (DIPTERA: CULICIDAE) AND THE RESISTANCE STATUS OF Aedes albopictus (FIELD STRAIN) AGAINST ORGANOPHOSPHATES IN PENANG, MALAYSIA. 2006 June. 44p.

[8] Christian Borgelt, Rudolf Kruse. Induction of Association Rules: Apriori Implementation. 6p.

[9] S. Baize et al., “emergence of zaire ebola virus disease in guinea — preliminary report,” New England Journal of Medicine, 16 April 2014. [10] Mitchell, Tom M. Machine Learning. New York: McGraw-Hill, 1997.