

Analysis of Relation between Aging and Telomere using Datamining – Apriori, Decision Tree, and Support Vector Machine(SVM)

Jongchan LEE, Taeseon YOON

Natural Science Course, Hankuk Academy of Foreign Studies, Yongin, Republic of Korea

Faculty of Computer: Hankuk Academy of Foreign Studies, Yongin, Republic of Korea

l63501@naver.com, tsyoon@hafs.hs.kr

Abstract— There have been a lot of researches about aging. Our study suggests a new research method using RTEL (Regulator of Telomere Elongation Helicase). Telomere is the region of repetitive DNA sequence at the end of chromosomes used as a buffer during the DNA replication. We made four sets (each set includes 3 mammals whose lifespan is almost similar) which have different lifespan range, and compared the RTEL DNA sequence among 4 sets in order to analyze how RTEL DNA sequence influence lifespan and aging. All of these studies were proved by Apriori, Decision Tree, SVM. Through this process, we found that the tendency of amino acid expression in each set is generally similar, but there are some amino acids which differentiate each other. Furthermore, we can support this research to Cancer since Telomere is receiving attention for main factor for finding the solution of cancer. Our research would be widely used as background knowledge for the research of Cancer and contribute to finding the relations between telomere and aging, lifespan in mammals.

Keywords— Apriori, Decision Tree, Support Vector Machine(SVM), Telomere, RTEL (Regulator of Telomere Elongation Helicase), Aging, Cancer, mammal

I. INTRODUCTION

Until 1970s, Scientists had believed that cells could divide eternally in the adequate circumstance. However, Hayflick discovered the number which cells can divide is limited, and different in each organism or organ. It is named ‘Hayflick limit’, which is found to be determined by ‘telomere’ at each end of a chromosome.

Telomere is a region of repetitive DNA sequence at the end of chromosomes, which protects end of chromosomes from the loss caused each cell division. So, telomere is getting shorter as cell division occurs. If telomere totally disappears, DNA having significant information will be lost and cell can develop a cancer. Therefore, at this moment, cell enters cellular old age(senescence), or begins cell self-destruction(apoptosis).

Telomere is generally being shortened but sometimes can be lengthened by telomerase reverse transcriptase, which acts

the task of adding repetitive DNA sequence to the end of the DNA. Generally, telomerase is active only in germ cells, stem cells in multicellular eukaryotic organism. However, in cancer cell, telomerase is activated much causing cancer cell not to enter cellular old age and to be immortal. So, telomere and telomerase are deeply involved with cancer.

For above reason, it is evident that cellular old age has direct relation to the telomere. However, the relation between overall aging and telomere is not apparent, and there have been many researches about aging.

In this research, we suggested a new research method using RTEL (Regulator of Telomere Elongation Helicase), which regulates telomere length diversity between mammalian species. We compared RTEL DNA sequence of mammals having different lifespan using Apriori, Decision Tree, SVM (Support Vector Machine) in order to analyze how RTEL DNA sequence influence aging and lifespan.

II. MATERIALS AND METHODS

A. DNA expression

DNA (Deoxyribonucleic acid) is a molecular carrying genetic information of all living organisms and many viruses. The sequence of DNA is called gene, which can influence the phenotype of an organism. In transcription, the codons referring 3 bases of gene are copied into messenger RNA(mRNA) by RNA polymerase. Then, each codon of mRNA is translated into specific amino acid which forms protein with the help of ribosome and transfer RNA(tRNA). Through this procedure, DNA sequence is expressed into protein.

B. Difference between mammals

We make four sets which have different lifespan range each other to analyze the relation between lifespan and DNA sequence. In each set, three mammals having similar lifespan are included for accurate comparison. Table 1 shows

mammals in each set and their approximate lifespan. This would be the background knowledge for comparing them.

TABLE 1. COMPARISON AMONG MAMMALS

Set 1	Life expectancy	Set 2	Life expectancy
Mus musculus	2-3	Ovis aries	10-12
Cricetulus griseus	2-3	Canis lupus familiaris	10-12
Ictidomys tridecemlineatus	3-4	Felis catus	13-15
Set 3	Life expectancy	Set 4	Life expectancy
Gorilla	35-40	Homo sapiens	66-70
Pan paniscus	35-40	Physeter catodon	68-72
Camelus dromedaries	40-45	Loxodonta Africana	60-70

C. Telomere

During chromosome replication, DNA polymerase, an enzyme which duplicates DNA, can't totally complete their duplication to the end of a chromosome because RNA primer should attach ahead of Okazaki fragment for synthesis of it in lagging strand. So, whenever the cell division ends, the end of each chromosome is shortened. Telomere plays a crucial role of disposal buffer at the end of chromosome

D. RTEL (Regulator of Telomere Elongation Helicase)

The genetic basis of telomere length heterogeneity among mammals is still not well known. Recently, a gene named RTEL (Regulator of Telomere Elongation Helicase) was identified to be essential element in species-specific telomere length regulation [1].

This gene encodes a DNA helicase that functions in the stability, protection and elongation of telomere and interacts with proteins in the shelterin complex which is known to protect telomere during DNA replication. Mutation in RTEL have been associated with Hoyerall-Hreidarsson syndrome and dyskeratosis congenita [2].

E. Apriori Algorithm

Apriori Algorithm is a common algorithm which shows the frequent item set mining and general rule of the given dataset [3]. This algorithm proceeds by two steps: the first step is identifying main frequent items in the dataset. The second step is to extend these frequent items to increasingly larger item sets as long as they show up sufficiently often in the data set [4]. Through the frequent item sets by Apriori Algorithm, general rules which can show feature of datasets are extracted. Apriori Algorithm uses a "bottom up" approach in which frequent subsets are extended once, which is known as candidate generation. Each group of candidates is tested continuously until no more successful expanded rules are found.

F. Decision Tree

Decision Tree is a common data mining tool widely used for leading influence, in order to generate a model which can predict the value of a main variable in accordance with several input variables. It contains nodes which require decisions for each questions about relative degree of influence on specific results [5].

In data mining, Decision Tree works as the mixture of computational and mathematical form, utilizing the progress of classification. Generally, the diagram gradually come in records of the form like this: $(X,Y)=(x_1,x_2,x_3,...x_k,\mu)$ [6].

Decision Tree uses entropy(H) to calculate the homogeneity of a sample.

$$H(S) = \sum (-P_i \log(P_i)) \quad (P: \text{given dataset})$$

G. Support Vector Machine(SVM)

Support Vector Machine (SVM), a very powerful algorithm primarily used for classification, regression analysis, and prediction, is a kernel-based margin classifier. Given a set of training examples, each belonged to one of two categories, SVM builds a non-probabilistic linear classifier model that allocates new examples into one category or the other bases on accumulated data given by previous instance. Input data is a representation of the examples as points in space, and SVM model maps the boundary so that the examples of the separate categories are divided by a definite gap that is as wide as possible [7]. As new examples are mapped into the space, experimenter is able to predict which side the examples would be categorized. Furthermore, in the situation when the data can not be categorized in linear form, SVMs can efficiently perform a non-linear classification, mapping their inputs into high dimensional feature space - "kernel trick".

H. Experiment

First, we gathered DNA sequence of RTEL of mammals in 2.2 from NCBI ("National Center for Biotechnology Information"). We performed experiment through Apriori algorithm, decision tree and SVM (Support Vector Machine). For each algorithm, we conducted experiment in 9 window, 13 window, and 17 window for each mammal. For Apriori algorithm and decision tree, we compared sets with each other by calculating the average of each set.

III. RESULTS

A. Apriori Algorithm

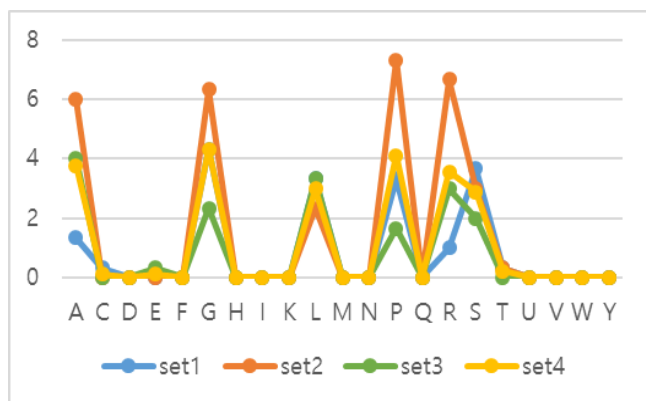


Figure 1. 9 window

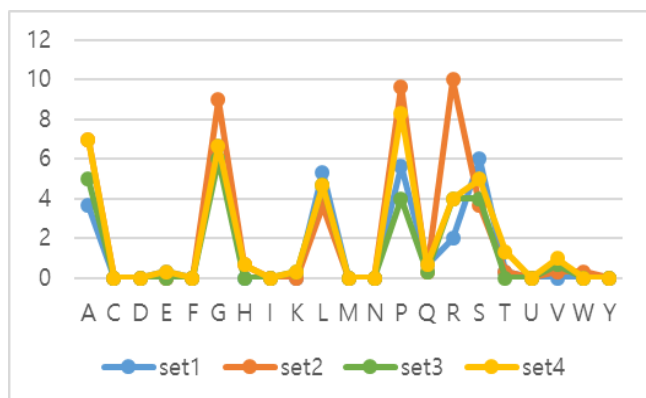


Figure 2. 13 window

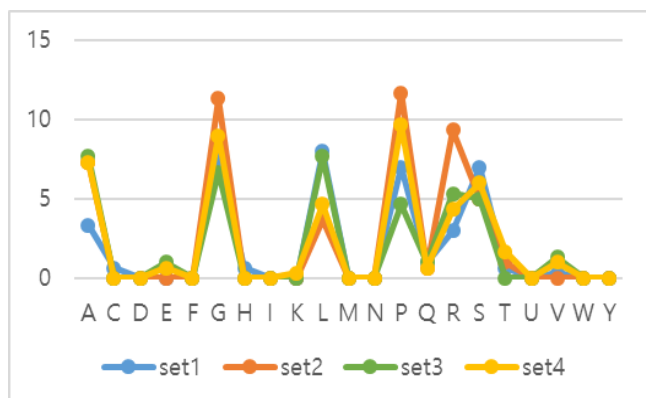


Figure 3. 17 window

A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y are names of amino acids. We counted the number of each amino acid and compared sets with each other in each window by calculating the average and drawing the graph.

According to Figure 1, Figure 2, Figure 3, the general tendency of amino acid expression is similar in each set. However, it is noticeable that there are some amino acids of which number considerably vary in each set. So, we assumed A, P, R, S are the important factors which differentiate each other.

B. Decision Tree

We extracted data with frequency rates 0.800 or higher. Among those rules, we counted the number of each amino acid and compared sets with each other by calculating the average, converting into a percentage, and drawing the graph.

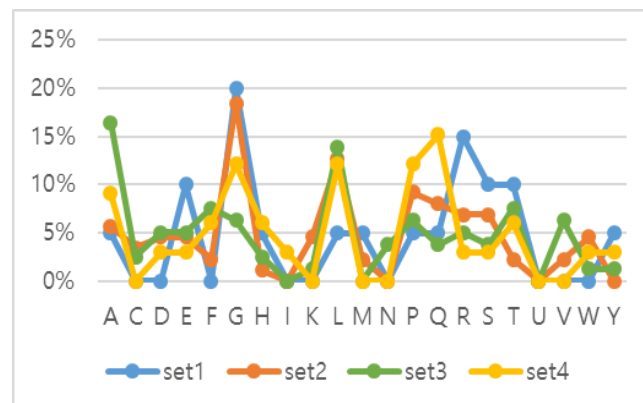


Figure 4. 9 window

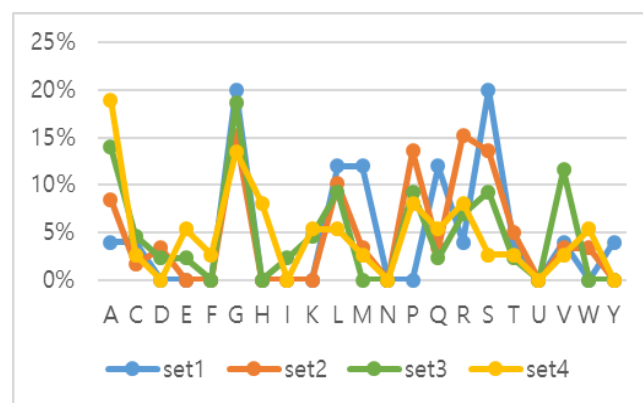


Figure 5. 13 window

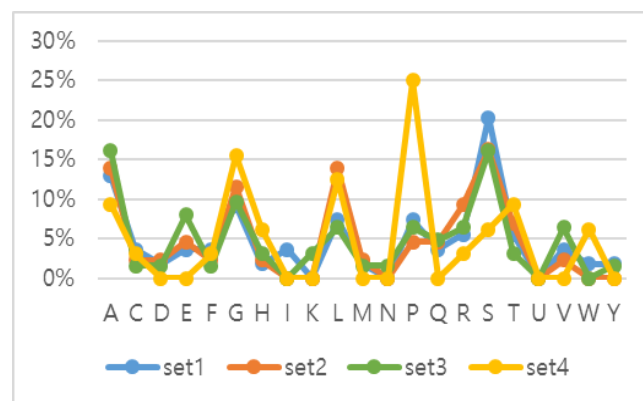


Figure 6. 17 window

According to Figure 4, Figure 5, Figure 6, the general tendency of amino acid expression is similar in each set but there are some amino acids of which number considerably vary in each set like the result of Apriori algorithm. We assumed A, P, Q, R, S are the important factors which differentiate each other.

C. Support Vector Machine (SVM)

We experimented with SVM multiclass program based on the 10-fold cross validation to classify the property of data. We used 4 criteria for evaluation: 1. Polynomial 2. RBF 3. Sigmoid 4. Normal.

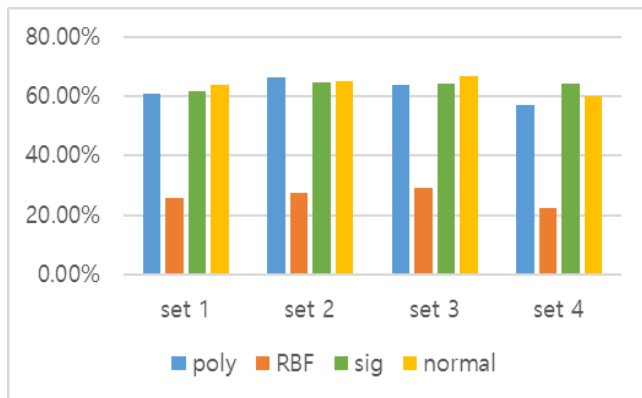


Figure 7. Average zero/one-error: 9 window

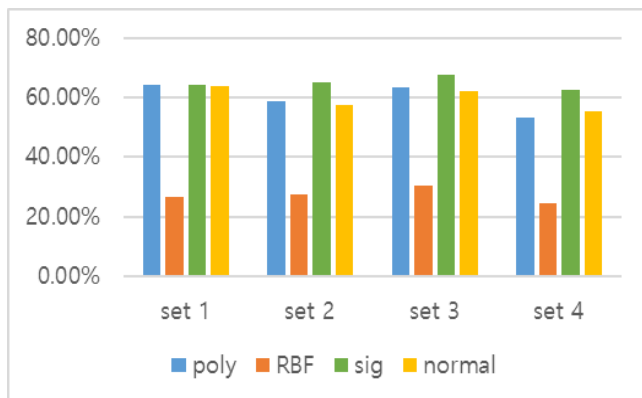


Figure 8. Average zero/one-error: 13 window

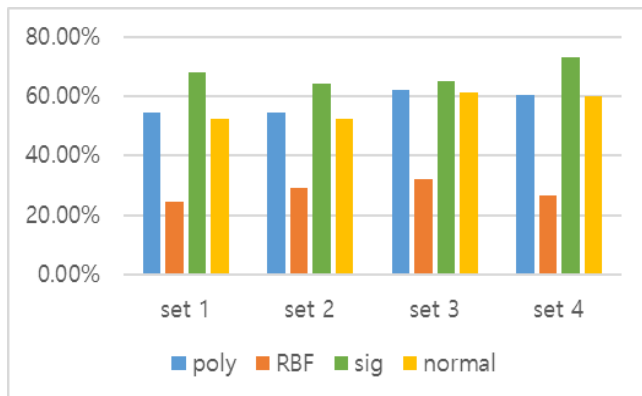


Figure 9. Average zero/one-error: 17 window

According to Figure 7, Figure 8, Figure 9, It's noticeable that three functions' (Polynomial, Sigmoid, Normal) zero/one-error value on test set turned to be high, approximately 60% on average, except RBF which is marked low, approximately 25%. Since RBF has a strong non-linear property, it generally

classifies data well. So, it shows relatively low zero/one-error value. A high zero/one-error value in three functions signifies that the data is difficult to classify.

IV. DISCUSSION

Telomere is the region of repetitive DNA sequence at the end of chromosomes used as a buffer during the DNA replication. It protects end of chromosomes from the loss caused each cell division and limits the number of cell division. Also, RTEL (Regulator of Telomere Elongation Helicase) is essential element in species-specific telomere length regulation. With this study, we analyzed relation between lifespan and RTEL through comparison of RTEL by using three kinds of algorithms: Apriori, Decision Tree, and SVM.

Apriori algorithm and Decision Tree showed the similar tendency of amino acid expression in each set, and we could find the molecular biological evidence of mammals. However, there are some amino acids which differentiate each other. Therefore, we concluded that A, P, R, S played a crucial role in determining aging of mammals as important factors in RTEL.

SVM showed that set1, set2, set3, set4 have high zero/one-error value in three functions. This result means that the data in each set is difficult to classify. Thus we concluded RTEL sequences of 3 mammals in each set are very similar. This fact demonstrates that our experiment method using average value of each set in comparison, and the result obtained from Apriori algorithm, Decision Tree are valid.

Our research is in value of relating the DNA sequence of RTEL and the aging using datamining. Also, by establishing each set including 3 mammals whose lifespan is similar, we could analyze relation between lifespan and RTEL in various aspects with more accuracy. In further research, we would like to increase the number of mammals in each set, and to observe more accurate tendency of amino acids. Also, we can support this research to Cancer since Telomere is receiving attention for main factor for finding the solution of cancer. We hope that our research will be widely used as background knowledge for the research of Cancer and contribute to finding the relations between telomere and aging, lifespan in mammals.

REFERENCES

- [1] Du,Z., Zhao,D., Zhao,Y., Wang,S., Gao,Y. and Li,N. (2007) Identification and characterization of bovine regulator of telomere length elongation helicase gene (RTEL): molecular cloning, expression distribution, splice variants and DNA methylation profile. BMC Mol. Biol., 8, 18.
- [2] "RTEL1 Regulator of Telomere Elongation Helicase 1 [Homo Sapiens (human)] - Gene - NCBI." National Center for Biotechnology Information. U.S. National Library of Medicine, n.d. Web. 05 Nov. 2016.
- [3] Leea, J.H., Ahna, S.H., Pyuna, S.M., Janga, E.J. and Yoon, T. (2014) Analysis of Malaria Inducing P. falciparum, P. ovale, and P. vivax through Apriori Algorithm and Decision Trees.

- [4] Borgelt, C. and Kruse, R. (2002) Induction of Association Rules: Apriori Implementation. Compstat. Physica-Verlag HD.+4
- [5] Lee, J.J. and Yoon, T. (2014) The New Approach on Fuzzy Decision Trees. International Journal of Fuzzy Logic Systems (IJFLS), 4.
- [6] Jang, S. P., K. H. Park, Y. L. Kim, H. N. Cho, and T. S. Yoon. "Comparison of H5N1, H5N8, and H3N2 Using Decision Tree and Apriori Algorithm." Journal of Biosciences and Medicines JBM03.06 (2015): 49-53. Web
- [7] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel Based Learning Methods, Cambridge University Press, 2000.



Jongchan Lee was born in Daejeon, Republic of Korea, in 1999 and currently attends the Natural Science Department of the Hankuk Academy of Foreign Studies. His major interest is in biology and computer science. He has been recently interested in bio-informatics and immunology and started his first research in 2015.



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, 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 in Yongin, Gyeonggi, where he is a Computer Science and Statistics Teacher. He was the recipient of the Best Teacher Award of the Science Conference, Gyeonggi-do, Korea, 2013.