2014 3rd International Conference on Environment Energy and Biotechnology
IPCBEE vol.70 (2014) © (2014) IACSIT Press, Singapore
DOI: 10.7763/IPCBEE. 2014. V70. 5

rRNA of Alphaproteobacteria Rickettsiales and mtDNA Pattern Analyzing by Decision Tree

Seung Jae Lim[+1], Taeseon Yoon[1]

[1]Natural Science Hankuk Academy of Foreign Studies

Abstract. Endosymbiosis is the most admitted theory of the cell evolution. There are some evidences which support endosymbiotic theory. Thanks to recent computer technologies, we can use computer intellectual based algorithms such as decision tree to analyze data set. In this paper, we were aspired to detect patterns and similarities among bacteria and mitochondria with employing decision tree. We've done 10-fold validation experiment to precisely detect rules of them. Furthermore, we've found some evidences of endosymbiosis and their evolutionary features.

Keywords: alphaproteobacteria rickettsiales, ribosomal RNA, mitochondria, decision tree.

1. Introduction

1.1. Endosymbiotic theory

After earth's born, the initial organisms occured. Time passed and they evolved into bacteria. The advent of prokaryote cells is considered as they triggered the beginning of planet's life. When three of prokaryote cells gathered, they formed eukaryotic cell. For instance, recently, escherichia coli bacteria is the representative organism included in prokaryotic organisms. The endosymbiotic theory insists that the aerobacteria were brought inside cells and through the long process of evolution, they settled down in cells. According to this theory, mitochondria and chloroplasts was free-living bacteria long time ago that were settled inside of another cell as an endosymbiont. There are some evidences which support this theory [1], [2].

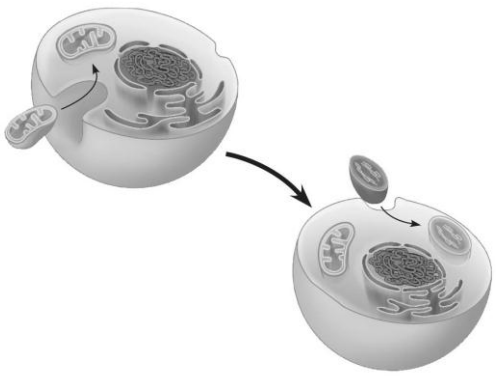

Fig. 1: Description of Endosymbiotic Theory [3]

1.2. Supporting evidences

As mentioned, there are some biological evidences exist. Mitochondria have their own DNA which is naked and circular. Fig 2 shows mitochondrial DNA. Circular DNA is the representative feature of the

+ Corresponding author. Tel.: + 82-31-704-4052; fax: +82-31-332-0042.
E-mail address: tonylim0930@gmail.com.

prokaryote cells while eukaryotic cells have strongly coiled DNA [5], [6]. Also, mitochondria has its own unique DNA, which is certainly differentiate from that of cell's [7]. So to speak, mitochondria DNA resembles that of prokaryote cells. Mitochondria has ribosomes that are similar to prokaryotes' which means that protein productions or enzymes might have similarities. Also, mitochondria has a double membrane and the inner membrane has proteins similar to prokaryotes. Plus, it is roughly the same size as bacteria and susceptible to the antibiotic chloramphenicol [8]-[10].

Lastly, it's noticeable that according to erstwhile studies, the genomes of mitochondria have similarities with that of the rickettsial bacteria.

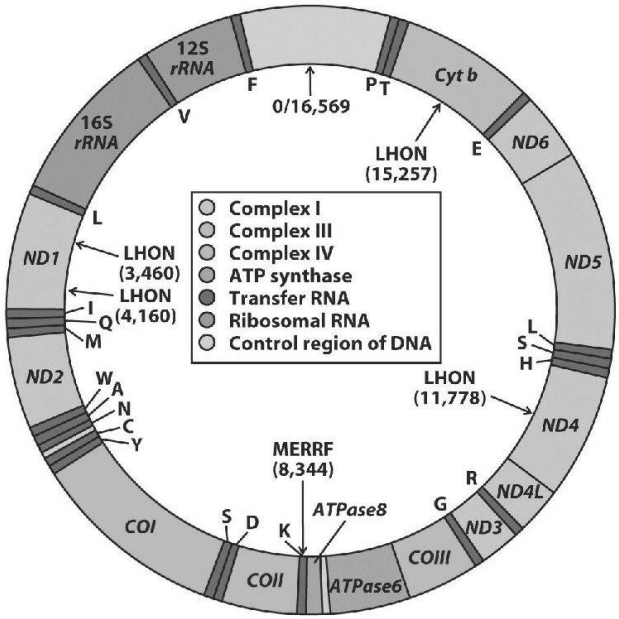

Fig. 2: Mitochondrial DNA [4]

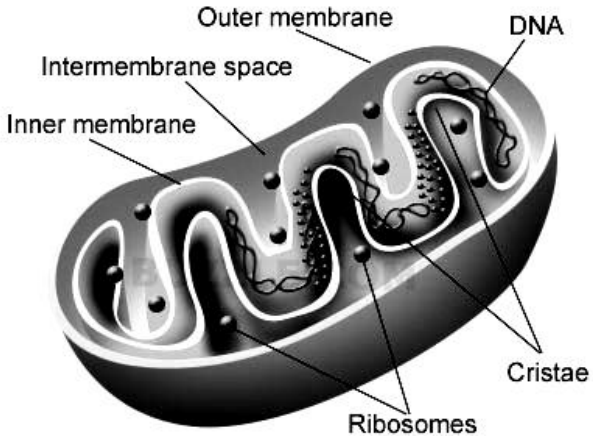

Fig. 3: Mitochondria Structural Features [11]

1.3. Decision tree

Decision tree is the representative analysis method of data mining [12]. Artificial intelligence, machine learning, and statistical analysis employ decision tree algorithm [13]. Normally, it is used to data classification and pattern recognizing [14]. However, it's impossible to predict. There are few algorithms of decision tree such as ID3, C4.5, and C5.0. C4.5 is revised version of ID3 and C5.0 is the upgraded version of C4.5 [15], [16]. Thus, explaining C4.5 is efficient. Among decision trees, entropy held a great role [17]. Entropy means the congestion of given data. So, if given data set contains many different classes, entropy appears to be high and if data set contains similar classes' records, entropy appears to below.

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad p_i = \frac{freq(C_i, S)}{|S|} \tag{1}$$

Equation (6) is the entropy equation. S means given data set, C = {C1, C2, ... , Ck} means the class set, freq(Ci,S) means the number of records which included in class Ci, and |S| is the number of data. The outcome of equation holds scope of 0 to 1. The highest congestion appears 1, and one class status appears 0. Decision tree algorithm classifies data set consequently high entropy to low entropy like tree's branch. When there are some more classes, S can be calculated with (7)

$$Entropy(a,b,c) = -\frac{a}{(a+b+c)}\log_2\frac{a}{(a+b+c)} - \frac{b}{(a+b+c)}\log_2\frac{b}{(a+b+c)} - \frac{c}{(a+b+c)}\log_2\frac{c}{(a+b+c)} \quad (2)$$

Additionally, Information gain plays important rule in decision tree algorithm. Information gain means that by choosing some attributes, which makes easily with classifying data set. Equation is given below.

$$Gain(A) = I(S_1, S_2, \ldots S_m) - E(A) \quad (3)$$

Equation (8) is the information gain equation. $I(S_1, S_2, \ldots S_m)$ means upper class's node which is subtraction of upper class's node and lower one and E(A) means the average entropy of divided nodes under attribute A. (7) is the equation that calculates information gain when it comes to attribute A and lager outcome represents larger information gain. Also, it means attribute A has great discrimination capacity.

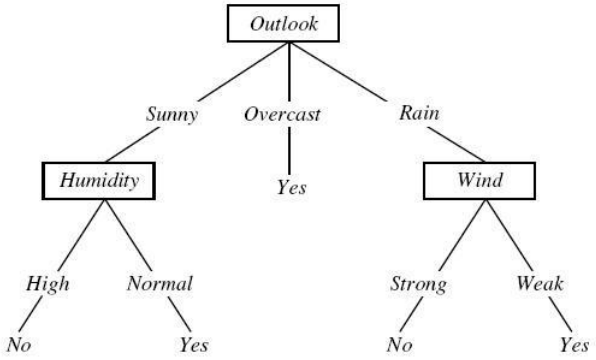

Fig. 4: A Simple Decision Tree [18]

2. Experiment Object

2.1. Rickettsia

Table I: Biological Classification of Rickettsia [19]

Kingdom	Bacteria
Phylum	Proteobacteria
Class	Alphaproteobacteria
Order	Rickettsiales
Family	Rickettsiaceae
Genus	Rickettsia
Species	-*Rickettsia aeschlimannii* Beati et al. 1997 -*Rickettsia africae* Kelly et al. 1996 -*Rickettsia akari* Huebner et al. 1946 (Approved Lists 1980) -*Rickettsia australis* Philip 1950 (Approved Lists 1980) -*Rickettsia bellii* Philip et al. 1983 -***Rickettsia canadensis* corrig. McKiel et**

al. 1967 (Approved Lists 1980)
-*Rickettsia conorii* Brumpt 1932 (Approved Lists 1980)
-*Rickettsia felis* Bouyer et al. 2001
-*Rickettsia heilongjiangensis* Fournier et al. 2006
-*Rickettsia helvetica* Beati et al. 1993
-*Rickettsia honei* Stenos et al. 1998
-*Rickettsia japonica* Uchida et al. 1992
-*Rickettsia massiliae* Beati and Raoult 1993
-*Rickettsia montanensis* corrig. (ex Lackman et al. 1965) Weiss and Moulder, 1984
-*Rickettsia parkeri* Lackman et al. 1965 (Approved Lists 1980)
-*Rickettsia peacockii* Niebylski et al. 1997
-***Rickettsia prowazekii* da Rocha-Lima 1916 (Approved Lists 1980)**
-*Rickettsia rhipicephali* (ex Burgdorfer et al. 1978) Weiss and Moulder, 1988
-*Rickettsia rickettsii* (Wolbach 1919) Brumpt 1922 (Approved Lists 1980)
-*Rickettsia sibirica* Zdrodovskii 1948 (Approved Lists 1980)
-*Rickettsia slovaca* Sekeyová et al. 1998
-*Rickettsia tamurae* Fournier et al. 2006
-*Rickettsia typhi* (Wolbach and Todd 1920) Philip 1943 (Approved Lists 1980)

Table I is the classification of rickettsia. Among some evidences of endosymbiotic theory, we've focused on genome resemblance which suggests mitochondria and bacteria have similar genomes. According to erstwhile studies of comparing of DNA between mitochondria and rickettsia showed that they have similarities among their genomic sequences. Especially, the genome sequence of *rickettsia prowazekii, rickettsia canadensis* and that of mitochondria presented large similarities by analyazing them with apriori and SVM [20]. In this paper, by analyzing them with decision tree, we'd like to find out specific rules and patterns among them. Also, analyzing them with different algorithm such as decision tree likely to come out some different results. We've focused on ribosomal RNA which plays role of organisms' protein (DNA, mRNA, tRNA, rRNA etc.) production. As mentioned, one of the evidence is known that mitochondrial ribosome and bacterial ribosome have resemblances with patterns and basic sequences. Thus, first of all, we've aspired to compare ribosomal RNA (rRNA) of *rickettsia prowazekii, rickettsia canadensis* and that of human mitochondria employing decision tree. All of them have 16S or 12S ribosomal RNA (rRNA) basic sequence data on NCBI (The National Center for Biotechnology Information) [21].

2.2. Mitochondrial DNA

Mitochondrial DNA (mtDNA or mDNA) is the DNA of mitochondria. It's known that mitochondria take charge in cell respiration. It converts chemical energy from food into adenosine triphosphate (ATP). According to Fig 2 in Introduction section, mitochondria has 12S and 16S rRNA. rRNA, which is ribosomal ribonucleic acid, plays an important role for protein synthesis in organisms [22]-[24].

3. Experiments

3.1. Decision tree

By employing decision tree(See 5.0) algorithm, 10 fold validation experiment held with 4 classes(canadensis, prowazekii, sapiens's 12S rRNA and sapiens's 16S rRNA). See 5.0 algorithm is based on C 5.0 and it can detect the rules more precisely.

3.2. Rule extraction

With the result of decision tree, we've selected high frequency data set in each window and each class. After this procedure, we've extracted amino acidic rules among classes.

4. Conclusion

4.1. Results

First of all, the result of experiment appears as a set of numerous numbers which represents 21 amino acids each. The number of data set appears 1 to 23 and 22, 23 have no meanings. The table Ⅱ given below shows Amino acids and their acronym. We've selected the data which shown relatively high frequency. Next, we've changed the number into amino acids.

Table Ⅱ: Amino acids & acronym [25]

Amino Acids	Acronym
Alanine	A
Cysteine	C
Aspartic acid	D
Glutamic acid	E
Phenylalanine	F
Glycine	G
Histidine	H
Isoleucine	I
Lysine	K
Leucine	L
Methionine	M
Asparagine	N
Proline	P
Glutamine	Q
Arginine	R
Serine	S
Threonine	T
Selenocysteine	U
Valine	V
Tryptophan	W
Tyrosine	Y

Table Ⅲ: Rule extraction under 9 window

	Rule	Frequency
Canadensis	pos2=I pos7=V	0.75
Prowazekii	pos4=K pos7=S	0.75

	pos2=T pos7=I	0.75
Sapiens 12S	pos3=G pos7=S	0.75
Sapiens 16S	pos7=D pos8=T	0.75
	pos4=L pos7=S	0.75
	pos2=S pos7=I	0.75

According to table Ⅲ, the rule frequency which overpassed 0.75 existed in little amount. Sapiens's 16S rRNA shown three rules which means that it has various features that differentiate it from other bacteria.

Table Ⅳ: Rule extraction under 13 window

	Rule	Frequency
Canadensis	pos9=G pos10=G	0.75
	pos8=G pos9=L	0.75
	pos6=S pos11=L	0.8
Prowazekii	pos9=G pos10=R	0.75
Sapiens 12S	Not extracted	0
Sapiens 16S	pos1=T pos9=A	0.75
	pos7=L pos11=K	0.75
	pos7=F pos11=K	0.75

In table Ⅳ, it's noticeable that Canadensis shown respectively increased number of rules. It is possible to infer that increased window condition contributed to such result. Also, it's noticeable that rule of the sapiens's 12S rRNA was not extracted. At the perspective of decision tree algorithm, we can assume that it's because of its similarities which makes it difficult to detect criteria of classification.

Table Ⅴ: Rule extraction under 17 window

	Rule	Frequency
Canadensis	pos4=R pos8=S	0.8
	pos5=L pos17=G	0.75
Prowazekii	pos4=D pos8=R	0.75
Sapiens 12S	Not extracted	0
Sapiens 16S	pos10=E pos17=I	0.75

According to table Ⅴ, the rule of sapiens's 12S rRNA was not extracted as same as the rule extraction under 13 window. Considering hypothesis that we've made in table Ⅳ, this same result proved that sapiens's 12S rRNA is included in other rules among canadensis, prowazekii and sapiens's 16S rRNA. However, we could firmly infer that it might be one of canadensis or prowazekii, because, sapiens's 12S and 16S rRNA have precisely different basic sequence due to they are both genomic sequence of one same organism (homo sapiens). This result can support endosymbiotic theory partially in that mt DNA's rRNA has similarities with bacteria (canadensis or prowazekii). However, Canadensis and prowazekii shown their own rules among all

window (9, 13, 17) condition which means that after endosymbiosis happened, they developed different amino acidic features while consequence of evolution.

4.2. Expectations

With the result of pattern analyzing with decision tree, we've found some evidences of endosymbiotic theory and rules of bacterial genomes. There are many evidences that support endosymbiosis. However, trial of finding evidences of endosymbiotic theory using computer algorithms was not common. However, in this paper, we used computer intellectual based algorithms such as decision tree (See 5.0) and by employing this algorithms, we found numerous unknown rules which can be used with classification task of bacteria and mitochondria. Plus, as the result revealed, our research found the evidences of endosymbiosis and its amino acidic rules and similarities. Using decision tree algorithm, we could easily detect rules and patterns of bacteria and mitochondria. Endosymbiosis is important theory which is related with evolution. The research we've made expected to contribute in understanding theories and its evidence.

5. References

[1] "Mitochondria Share an Ancestor With SAR11, a Globally Significant Marine Microbe". ScienceDaily. July 25, 2011. Retrieved 2011-07-26

[2] J. Cameron Thrash et al. (2011). "Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade." Scientific Reports.

[3] Endosymbiosis, http://hermansonhbiology.wordpress.com/2011/11/03/(website)

[4] Mitochondrial DNA http://chimerasthebooks.blogspot.kr/2011/12/another-genetic-puzzle-why-is. html(website)

[5] F. J. borra, H. Kimura, P. R. Cook. "The functional organization of mitochondrial genomes in human cells." *BMC Biol.* **2**: 9. 2004.

[6] B. Sykes, (10 September 2003). "Mitochondrial DNA and human history." *The Human Genome*. Wellcome Trust. Retrieved 5 February 2012

[7] "Mitochondrial DNA: The Eve Gene." *Bradshaw Foundation*. Bradshaw Foundation. Retrieved 5 November 2012

[8] J. Kimball, 2010. Kimball's Biology Pages. Accessed October 13, 2010. An online open source biology text by Harvard professor, and author of a general biology text, John W. Kimball.

[9] J. Reece, A. Lisa Urry, L. Michael Cain, A. Steven Wasserman, V. Peter Minorsky, B. Robert Jackson, 2010. Campbell Biology. 9th Edition Benjamin Cummings; 9th Ed. (October 7, 2010)

[10] P. Raven, George Johnson, Kenneth Mason, Jonathan Losos, Susan Singer, 2010. Biology. McGraw-Hill 9th Ed. (January 14, 2010)

[11] Mitochondrial structure, http://www.buzzle.com/articles/mitochondria-structure-and-functions.html (website)

[12] J. R, QUINLAN, "Induction of Decision Trees", 1986 Kluwer Academic Publishers, Boston, Machine Learning 1: pp.81-106, 1986

[13] Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8

[14] Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711

[15] J. J. Rodriguez, and L. I. Kuncheva, and C. J. Alonso, (2006), Rotation forest: A new classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10):1619-1630

[16] G. V. Kass, (1980). "An exploratory technique for investigating large quantities of categorical data". *Applied Statistics* 29 (2): 119–127

[17] C. Barros, Rodrigo, M. P. Basgalupp, A. C. P. L. F. Carvalho, Freitas, Alex A. (2011). A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 42, n. 3, p. 291-312, May 2012.

[18] J.R, QUINLAN, "Induction of Decision Trees", 1986 Kluwer Academic Publishers, Boston, Machine Learning 1:

pp.81-106, 1986

[19] Encyclopedia Of Life, http://eol.org/pages/3349/overview

[20] Seung Jae Lim, Shin Hyo Bang, Dae Seop Kim and Tae Seon Yoon, "rRNA of Alphaproteobacteria Rickettsiales and mtDNA Pattern Analyzing with Apriori & SVM"**,** Lecture Note Computer Science (LNCS), Biologically Inspired Techniques for Data Mining (BDM'14), 13-16 May, 2014, Tainan, Taiwan

[21] The National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/(website)

[22] M. S. Meusel, R. F. Moritz (1993). "Transfer of paternal mitochondrial DNA during fertilization of honeybee (Apis mellifera L.) eggs". *Curr. Genet.* **24** (6): 539–43

[23] U. Gyllensten, D. Wharton, A. Josefsson, A. C. Wilson (1991). "Paternal inheritance of mitochondrial DNA in mice". *Nature* **352** (6332): 255–7

[24] "Genetic Genealogy". Ramsdale.org. 2003-05-19.doi:10.1371/journal.pbio.1000285. Retrieved 2012-07-14.

[25] Amino Acids, http://www.wikipedia.org/(website)