

Computational Analysis of Chloroplast DNA and Synechococcus by Apriori and Decision Tree

Kyuhee Jo*, Hyemin Lee**, Taeseon Yoon**

*International science Department, Hankuk Academy of Foreign Studies, 449 854 Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, South Korea

**Natural Science Department, Hankuk Academy of Foreign Studies, 449 854 Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, South Korea

kyuhee0622@hafs.hs.kr, stephanie1904@hafs.hs.kr, tsyoon@hafs.hs.kr

Abstract— This research aims to discover validate the endosymbiotic theory of chloroplast by comparing chloroplast DNA and cyanobacterial DNA using computational methods such as apriori and decision tree. We compared the nucleotide sequences of four plants that each represent four evolutionary classes, bryophytes, pterophyts, gymnosperms, and angiosperms, and Synechococcus DNA. The rules extracted from five experiment object commonly shown abundance in involution of aminoacid leucine and serine. 4 to 11 rules were shared between each chloroplast DNA and the Synechococcus DNA, supporting endosymbiotic theory. Also, the rules extracted from decision tree shown similar patterns, implicating the similar evolutionary steps of the chloroplast from cyanobacteria for different classifications.

Keywords— Chloroplast DNA, Endosymbiotic theory, Synechococcus, Decision tree, Apriori.

I. INTRODUCTION

This document is a template. Endosymbiosis is the most convincing theory of the cell evolution. According to the preceding research, it was already revealed that bacteria and mitochondria have similarities by analyzing data set in computer intellectual based algorithm, apriori and SVM. [8] In this paper, by showing the decision tree data of chloroplast which is another evidence of endosymbiosis, this study aims to generalize the apriori algorithm data of endosymbiosis. In addition, we examine the result by applying it another algorithm, decision tree.

II. RELATED RESEARCH

A. Endosymbiotic Theory

Endosymbiosis theory explains that by prokaryotic cells which have different functions living in the other cell as intracellular symbiont, eukaryote was created. According to this theory, prokaryotes which bred independently turns to live together in one cell with symbiotic relationship, and part of them enter another protista and turned to organelles which have different functions each. The theory was created to explain the origin of organelles which have own DNA apart and synthetic enzyme independently such as mitochondria and chlorophyll. [1,9]



Figure 1. Description of endosymbiotic theory

B. Chloroplast DNA

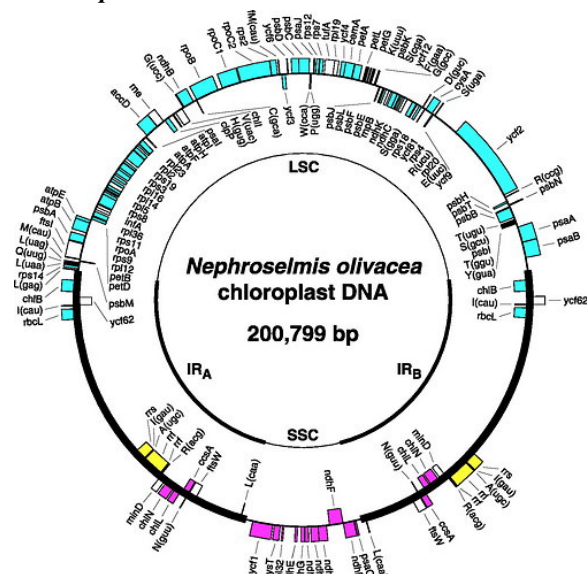


Figure 2. Typical example of chloroplast DNA[20]

Since its first discovery, the usage of molecular methods led to identification of general characteristics of chloroplast DNA, or ctDNA. Similar to typical prokaryotes' DNA, Chloroplasts contain circular DNAs, normally in size of 120,000-170,000 base pairs long. [5, 19] Its mass is around 80-130 million daltons, with a contour length near 30-60 micrometers. [3] While the nuclear DNA encode some proteins in the chloroplasts, most of the proteins are encoded by the ctDNA.

[12] Some studies suggest ctDNAs' possession of around 100 potential genes. [17]

C. Apriori Algorithm

Apriori is a rather simple algorithm first developed for the purpose of analyzing databases regarding financial transactions. [16] By extracting items that are frequent in a certain database, it proceeds to bigger sets until the item does not appear in a significantly large number. As a result, Apriori identifies the frequent item sets, that represents the general tendency of the whole data set. This then can be used to find out the association rule, where existence of one item can guarantee that of another.

Most notable application of Apriori is in the case of market basket analysis. Association rules and interrelations of different market products are analyzed by the usage of Apriori, and often these discoveries are later employed in creating marketing strategies. [10] Extended applications include sequence analysis of different genomes, which will be the case of this research.

D. Decision Tree

Decision tree is an algorithm that support decision making—more specifically classifying and predicting—characterized with its model of unique tree structure. The process of creating a decision tree progresses by (1) training (2) creation of model (3) application. The training of a decision tree requires sets of training data composed of the value of each column, representing a certain attribute, and the result, which is the final goal of prediction. This composition is aimed to meet the the algorithm's way of making prediction by evaluating the characteristics of a certain data set. Then, out of the numerous columns, the decision tree extracts the best decision rule [14], in other words selecting the best split, according to the calculation of information gain. Information gain is calculated by comparing the impurity of parent node (before splitting) and child node (after splitting). Examples of impurity measures are as follows:

$$\text{Entropy}(t) = - \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t), [15]$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2, [2]$$

where $p(i|t)$ is the ratio of records corresponding to class i at a given a node t , c is the number of classes, and $0 \log_2 0 = 0$ in entropy calculations. And the information gain, based on the impurity, is defined as

$$I(\text{parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j),$$

where I is the impurity measure, N is the total records at the parent node, k is the number of attribute values (columns), and $\frac{N(v_j)}{N}$ is the number of records at the child node, v_j . The bigger the information gain is the better a certain decision rule

is at classifying the data. Hence, the decision tree model selects the split with greatest information gain.

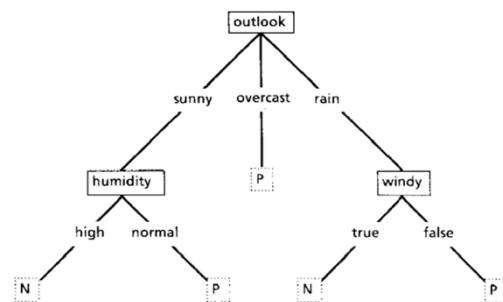


Figure 3. Example of a decision tree [13]

The biggest advantage of using decision tree is that one can easily observe the process of making decision, unlike other popular algorithms such as neural network system or support vector machine where the model is presented mainly as a set of number.

III. EXPERIMENT OBJECT

A. Chloroplast DNA

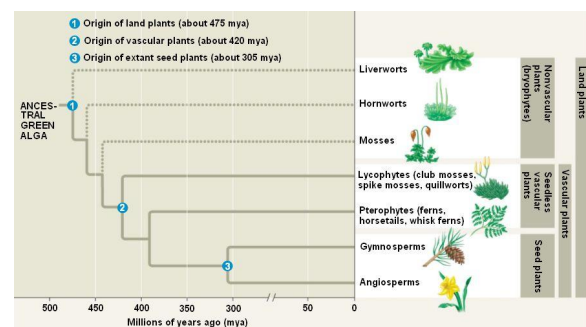


Figure 4. Brief summary of plant evolution

Starting from the ancestor of Kingdom Plantae, which endosymbiotic theory suggests as the fusion of a prokaryotic cell and cyanobacteria, plants have evolved in many ways. [18] (Figure 4) Since the purpose of the research is not only to examine the authenticity of endosymbiotic theory but also to identify the evolutionary relationships of plants by its similarity to its ancestor, cyanobacteria, four plants' ctDNAs were chosen as a representative of four evolutionary classes, which are bryophytes, pterophytes, gymnosperms, and angiosperms. The full sequence of chloroplast of *Ceratophyllum demersum* was chosen as a representative of bryophytes, *Osmunda cinnamomeum* for pterophytes, *Prunus yedoensis* for angiosperms and *Cephalotaxus oliveri* for gymnosperms. [11, 7, 4, 21] Bryophytes were differentiated primarily, then pterophytes, and gymnosperms, and angiosperms in that order.

1) *Ceratophyllum demersum*: *Ceratophyllum demersum* is a plant commonly known as hornwort, classified as order Ceratophyllales, family Ceratophyllaceae, and genus *Ceratophyllum*. Being an aquatic plant, it dwells in fresh water

such as lake and pond, widely spread in almost all region except Antarctica. The plant is monocious as well.

2) ***Osmundastrum cinnamomeum***: *Osmundastrum*, commonly known as cinnamon fern, is in genus *Osmundastrum* of family *Osmundaceae* of order *Osmundales*. It is a terrestrial plant prominent in the Americas and eastern Asia. Its habitat is usually swamp areas.

3) ***Prunus yedoensis***: *Ceratophyllum prunus yedoensis*, also known as Yohshino cherry, is created by hybridizing *Prunus speciosa* and *Prunus pendula* f. *ascendens*, both naturally and artificially. It is classified as order *Rosales*, family *Rosaceae*, and genus *Prunus*. It is a deciduous tree that grows up to 5 to 12 m. Its origin is in Japan and is introduced in Europe and North America as well.

4) ***Cephalotaxus oliveri***: *Cephalotaxus oliveri*, in class *Pinopsida*, order *Pinales*, family *Cephalotaxaceae*, and genus *Cephalotaxus* is a small tree or shrub that grows up to 4 meters. Its native ranges from China to other eastern nations such as Thailand, Laos, Vietnam and eastern India.

B. Cyanobacteria

Cyanobacteria is a phylum under domain bacteria that is capable of photosynthesis. Out of numerous species of the phylum, the full sequence of *Synechococcus* will be used for the experiment. [22] *Synechococcus* is in order *Synechococcales*, family *Synechococcaceae*, and Genus *Synechococcus*. Past studies show the close relationship between plant chloroplast and *Synechococcus* in terms of the protein and DNA structure. [6] Bearing this in mind, further comparing *Synechococcus* with different ctDNA of different species of plants would bring substantial proofs to support the endosymbiotic theory of chloroplast.

IV. EXPERIMENT METHOD

A. Apriori

Apriori algorithm extracts rules that characterizes the experiment object. Using Apriori algorithm, we identified unique patterns for the chloroplast DNA of four plants—*Ceratophyllum demersum*, *Osmundastrum cinnamomeum*, *Prunus yedoensis*, *Cephalotaxus oliveri*, and *Synechococcus*. For each objects, experiments were done three times under 13, 17, 19 window. Then, by comparing the rules of 4 plants with *Synechococcus*, we identified rules that apply to the plant and *Synechococcus* in common.

B. Decision Tree

Decision Tree algorithm operates by comparing two different sequences, and identifying rules to distinguish the two objects. Since the purpose of this result to compare cyanobacteria and four experiments were done in comparison between *Synechococcus* and the chloroplast DNA of four plants. Since the *Synechococcus* has sequence of length about two folds larger than other objects, parts for analysis were randomly selected. Like Apriori, three different trials were done for each pair of comparison under 13, 17, 19 window. The

training and testing were done 10times (10 folds) for each trial in order to increase authenticity. Once again, the frequencies of amino acids introduced in the extracted rule were analyzed.

V. RESULTS

For the comfort of organization, we used acronyms to represent each amino acids, which are shown below.

TABLE 1. ACRONYMS FOR AMINO ACIDS

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

A. Apriori

From the nucleotide sequence of *Synechococcus*, Apriori extracted 1 rule under 13 windows, 5 rules under 17 windows, 13 rules under 19 windows. Majority of the rules involved amino acid leucine. Serine also played role in few rules and arginine in one rule. All rules extracted is annotated in the following table.

TABLE 2. RULES EXTRACTED FROM SYNECHOCOCCUS UNDER 13 WINDOW

rule	frequency
amino9=L	359

TABLE 3. RULES EXTRACTED FROM SYNECHOCOCCUS UNDER 17 WINDOW

rule	frequency
amino14=L	264
amino6=L	264
Amino5=L	263
amino12=L	260
amino2=L	246

TABLE 4. RULES EXTRACTED FROM SYNECHOCOCCUS UNDER 19 WINDOW

rule	frequency
amino7=R	242
amino3=L	240
amino15=L	239
amino13=L	239
amino14=S	239
amino8=S	234
amino6=L	232
amino18=L	232
amino2=L	231
amino9=L	231
amino1=L	230
amino5=L	230
amino10=S	230

1) *Ceratophyllum demersum*: Apriori extracted 7 rules under 13 window, 11 rules under 17 window, 13 rules under 19 window. 6 rules involved serine and 1 rule involved leucine under 13 window, 8 rules of serine and 3 rules of leucine under 17 window, 8 rules of serine and 5 rules under leucine under 19 window. Rules in common with *Synechococcus* are shown in the following table.

TABLE 5. COMMON RULES BETWEEN SYNECHOCOCUS AND CERATOPHYLLUM DEMERSUM

window	rule(frequency)
13 window	No same rule
17 window	amino 14=L (316)
19 window	amino14=S (277) amino8=S (298) amino10=L (279)

2) *Osmundastrum cinnamomeum*: Apriori extracted 20 rules under 13 window, 22 rules under 17 window, 25 rules under 19 window. 13 rules involved serine and 7 rule involved leucine under 13 window, 15 rules of serine and 7 rules of leucine under 17 window, 17 rules of serine, 7 rules under leucine and notably one rule of Isoleucine under 19 window. Following table shows rules shared with *Synechococcus*.

TABLE 6. COMMON RULES BETWEEN SYNECHOCOCUS AND OSMUNDASTRUM CINNAMOMEUM

window	rule(frequency)
13 window	amino9=L 372
17 window	amino15=L 283 amino12=L 286 amino2=L 301
19 window	amino13=L 258 amino14=S 287 amino8=S 263 amino6=L 266 amino18=L 258 amino2=L 266 amino10=S 257

3) *Prunus yedoensis*: Apriori extracted 9 rules under 13 window, 13 rules under 17 window, 11 rules under 19 window. The majority of rules involved leucine. 6 rules involved leucine and 3 rule involved serine under 13 window, 8 rules of leucine and 5 rules of serine under 17 window, 8 rules of leucine, 3 rules under serine and notably one rule of Isoleucine under 19 window. Following table shows common rules.

TABLE 7. COMMON RULES BETWEEN SYNECHOCOCUS AND PRUNUS YEDOENSIS

window	rule(frequency)
13 window	No same rule
17 window	amino6=L (332)
19 window	amino13=L (287) amino15=L (280) amino6=L (283) amino18=L (312) amino1=L (286)

4) *Cephalotaxus oliveri*: Apriori extracted 12 rules under 13 window, 17 rules under 17 window, and 15 rules under 19 window. Notably, all rules under 13 window involved leucine. For each 17 window and 19 window, except one rule that involved Isoleucine, all rules involved leucine. Rules in common are shown in the following table.

TABLE 8. COMMON RULES BETWEEN SYNECHOCOCUS AND CEPHALOTAXUS OLIVERI

window	rule(frequency)
13 window	amino9=L (358)
17 window	amino14=L (295) amino5=L (283) amino12=L (273) amino2=L (288)
19 window	amino13=L (282) amino15=L (249) amino18=L (270) amino2=L (241) amino9=L (253) amino1=L (283)

B. Decision Tree

The rules extracted from decision tree by comparing the nucleotide sequences of two different experimental objects are that used to discriminate the two sequences. Hence, the rules extracted from comparing four chloroplast DNA and the *Synechococcus* bacteria represents the difference between chloroplast and the cyanobacteria, implicating how the chloroplasts evolved from cyanobacteria.

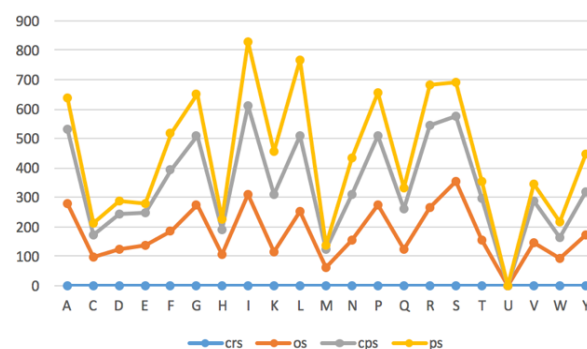


Figure 4. Decision tree results Under 19 window

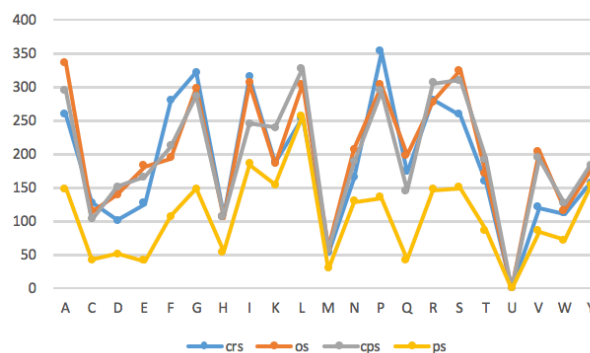


Figure 4. Decision tree results Under 17 window

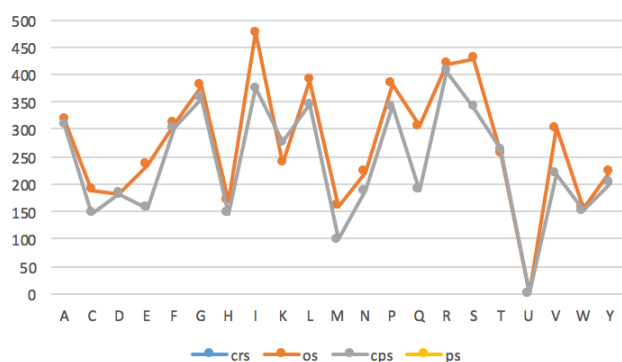


Figure 4. Decision tree results Under 13 window

The three graphs above explain the number of each amino acid involved in the extracted rules which represents the contribution of each amino acid in the evolutionary process. Crs is an abbreviation for *Ceratophyllum demersum*, os for *Osmundastrum cinnamomeum*, cps for *Cephalotaxus oliveri* and ps for *Prunus yedoensis*. Uniquely, the overall patterns of the graphs for each chloroplast DNA were very similar. Also, no rules were extracted for crs under 19 window and crs and ps under 13 window.

V. CONCLUSION

From the analysis of chloroplast DNAs of four different plants that each represents bryophytes, pterophytes, gymnosperms, and angiosperms, and the DNA of a cyanobacteria, *Synechococcus*, we were able to obtain evidences that support endosymbiotic theory. Firstly, the rules extracted from *Synechococcus* by apriori involved mainly two types of amino acids, leucine and serine. This is similar for the cases of the rules extracted from the chloroplast DNA by apriori, also consisted of mainly leucine and serine. Secondly, the bacteria and the plants share multiple rules extracted by apriori. 4 common rules were extracted from *Synechococcus* with *Ceratophyllum demersum*, 11 common rules with *Osmundastrum cinnamomeum*, 6 common rules with *Prunus yedoensis*, and 11 common rules with *Cephalotaxus oliveri*. These evidences support that chloroplast evolved from cyanobacteria. On the other hand, the experiments using decision tree suggested that the chloroplast of plants in different evolutionary classes, yet bonded under same kingdom, plantae, evolved from the cyanobacteria in similar way. The rules that distinguished four plants from *Synechococcus* involved similar proportions of amino acids, showing how the differences are similar. Like this, by using computational methods, our experiment successfully performed evidences of endosymbiotic theory from a very unique perspective.

Still, some questions remained unanswered. Firstly, the role of abundantly introduced amino acids, most notably leucine and serine, should be identified. Secondly, this research failed to find evidence that supports the plant phylogenetic tree regarding the time period bryophytes, pterophytes, gymnosperms, and angiosperms were introduced. Thirdly, the reason why different rules were extracted in different window

should be studied. Thirdly, new analyses can arise by considering the position of amino acids in the decision trees rule along with the number of rules involving each amino acids, which this research only focused on.

REFERENCES

- [1] Baum, D. A., & Baum, B. (2014). An inside-out origin for the eukaryotic cell. *BMC Biology* *BMC Biol*, 12(1). doi:10.1186/s12915-014-0076-2
- [2] 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
- [3] Burgess, Jeremy (1989). *An introduction to plant cell development*. Cambridge: Cambridge university press. p. 62. ISBN 0-521-31611-1.
- [4] Cho, M., Cho, C. H., Kim, S. Y., Yoon, H. S., & Kim, S. (2015). Complete chloroplast genome of *Prunus yedoensis* Matsum. (Rosaceae), wild and endemic flowering cherry on Jeju Island, Korea. *Mitochondrial DNA Part A*, 27(5), 3652-3654. doi:10.3109/19401736.2015.1079840
- [5] Clegg, M. T.; Gaut, BS; Learn Jr, GH; Morton, BR (1994). "Rates and Patterns of Chloroplast DNA Evolution". *Proceedings of the National Academy of Sciences*. 91 (15): 6795–801. Bibcode:1994PNAS...91.6795C. doi:10.1073/pnas.91.15.6795. PMC 44285. PMID 8041699.
- [6] Cozens, A., & Walker, J. (1987). The organization and sequence of the genes for ATP synthase subunits in the cyanobacterium *Synechococcus* 6301. *Journal of Molecular Biology*, 194(3), 359-383. doi:10.1016/0022-2836(87)90667-x
- [7] Kim, H. T., Chung, M. G., & Kim, K. (2014). Chloroplast Genome Evolution in Early Diverged Leptosporangiate Ferns. *Molecules and Cells*, 37(5), 372-382. doi:10.14348/molcells.2014.2296
- [8] Lim, S. J., Bang, S. H., Kim, D. S., & Yoon, T. (2014). rRNA of Alphaproteobacteria Rickettsiales and mtDNA Pattern Analyzing with Apriori & SVM. *Lecture Notes in Computer Science Trends and Applications in Knowledge Discovery and Data Mining*, 112-122. doi:10.1007/978-3-319-13186-3_11
- [9] Lim, S. J., Bang, S. H., Kim, D. S., & Yoon, T. (2014). RRNA of Alphaproteobacteria Rickettsiales and mtDNA Pattern Analyzing with Apriori & SVM. *Lecture Notes in Computer Science Trends and Applications in Knowledge Discovery and Data Mining*, 112-122. doi:10.1007/978-3-319-13186-3_11
- [10] Liu, Y., & Guan, Y. (2008). FP-Growth Algorithm for Application in Research of Market Basket Analysis. 2008 IEEE International Conference on Computational Cybernetics. doi:10.1109/iccocyb.2008.4721419
- [11] Moore, M. J., Bell, C. D., Soltis, P. S., & Soltis, D. E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences*, 104(49), 19363-19368. doi:10.1073/pnas.070807210
- [12] Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Ozeki, H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*, 322(6079), 572-574. doi:10.1038/322572a0
- [13] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106. doi:10.1007/bf00116251
- [14] Quinlan, J. R. (1987). "Simplifying decision trees". *International Journal of Man-Machine Studies*. 27 (3): 221. doi:10.1016/S0020-7373(87)80053-6.
- [15] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993
- [16] Rakesh, A., Ramakrishnan, S. (1994). Fast algorithms for mining association rules. *Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases*, 487-499.
- [17] Ravi, V., Khurana, J. P., Tyagi, A. K., & Khurana, P. (2006). The chloroplast genome of mulberry: Complete nucleotide sequence, gene organization and comparative analysis. *Tree Genetics & Genomes*, 3(1), 49-59. doi:10.1007/s12955-006-0051-3
- [18] Reece, J. B., & Campbell, N. A. (2011). *Campbell biology*. Boston: Benjamin Cummings / Pearson. 604-606.
- [19] Shaw, J.; Lickey, E. B.; Schilling, E. E.; Small, R. L. (2007). "Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III". *American Journal of Botany*. 94 (3): 275 – 88. doi:10.3732/ajb.94.3.275. PMID 21636401.
- [20] Turmel, M., Otis, C., & Lemieux, C. (1999). The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: Insights into the

architecture of ancestral chloroplast genomes. Proceedings of the National Academy of Sciences, 96(18), 10248-10253. doi:10.1073/pnas.96.18.10248
[21] Yi, X., Gao, L., Wang, B., Su, Y., & Wang, T. (2013). The Complete Chloroplast Genome Sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): Evolutionary Comparison of *Cephalotaxus* Chloroplast DNAs and Insights into the Loss of Inverted Repeat Copies in Gymnosperms. *Genome Biology and Evolution*, 5(4), 688-698. doi:10.1093/gbe/evt042



Kyuhee Jo was born in 2000, in Seoul, Korea. She is currently a student in department of international studies of Hankuk Academy of Foreign Studies. She is currently studying computational biology.

She is interested in algorithms that can be used to analyze nucleotide sequences, and its application in diverse fields of biology.

Hyemin Lee was born in 2000. She is currently a student in natural science major of Hankuk Academy of Foreign Studies. She is studying computational biology and bioinformatics.

She is interested in biology, in particular bioengineering. She is studying computer-based algorithms and its application at the field of biology.



Taseon Yoon was born in Seoul, Korea, in 1982. He was Ph.D. Candidate degree in Computer education from 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 Anssan University, as a Adjunct professor. Since December 2004, he has been with the Hankuk Academy of Foreign Studies, where he was a Computer Science and Statistics Teacher.

He was the recipient of the Best Teacher Award of the Science Conference, Gyeonggi-do, Korea, 2013.