Analysis on Mannose-binding Lectin as a Treatment of *Helicobacter pylori* by Using Data Mining

Seojun Kim*, Wonjong Lee*, Juho Jung*, and Taeseon Yoon**

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

**Computer Science and Engineering, Korea University, Seoul, Republic of Korea

seojunkim@hafs.hs.kr, wonjong720@gmail.com, jhjung@hafs.hs.kr, tsyoon@hafs.hs.kr

***Abstract* - As a critical role in overall human body reactions to foreign organisms, innate immune system, especially Mannose-binding lectin (MBL), has worked for the preservation of life. Since targeted therapy on bacterial infection using innate immune system has been researched for destroying pathogens without harming ourselves, MBL could be used for the targeted therapy. Based on three algorithms; Decision Tree Algorithm, Apriori Algorithm and Support Vector Machine, analysis on chemical bond formation by comparing the similarities between two proteins which have direct relevance with mannose could suggest the potential of utilizing proteins of MBL for targeting foreign factors. According to the results, Helicobacter pylori and Homo sapiens showed distinguishable features but indicated a few common factors. We could improve the targeting treatments by considering immunological approach using MBL; to analyze the possibility for forming chemical bond between human MBL and mannose of Helicobacter pylori.**

***Keywords*- data mining, immune system, mannose-binding lectin, C3b, helicobacter pylori, bioinformatics**

I. INTRODUCTION

As a critical role in overall human body reactions to foreign organisms, innate immune system that is mostly consisted of complement system, professional phagocytes, and natural killer cells, has worked for the preservation of human life. The immune system has been using the most effective way for destroying pathogens, derived from the accumulated experiences of natural selection.

It is showed that innate immune system initiates identical responses to diverse pathogens. For instance, these characteristics of immune system could be observed in the activation pathways of C3b complement system. Especially, mannose-binding lectins (MBL), one of the main factors of complement system, can also be activated by general pathogens. However, unlike C3b complement system, MBL initiate specific mechanism which targets foreign mannose[1].

Since targeted therapy on bacterial infection has been consistently researched for its efficiency of destroying pathogens without harming our own body cells[2], MBL could be an improvement for the targeted therapy. Using its ability to eliminate pathogens by forming chemical bond between MBL and foreign mannose, researching the possibility of destroying bacteria, represented as *Helicobacter pylori*, could be considered.

II. MATERIALS AND METHODS

A. Complement System

1. Alternative Pathway

Alternative pathway of complement system is one of the pathways of complement system activation. C3b proteins are considered as a main factor of this mechanism, while consisting other proteins that cooperate with each other to prevent bacterial infection.

C3 is known as an important protein in this process, and specifically, the most spontaneously well-disintegrated protein is C3b. This shows high rate of reaction to the surface of bacteria. C3b can be activated with protein D. This mutates the structure of C3b and induces activation of other C3b protein units, which is considered as a positive feedback. Then, the staircase procedure is followed, which consists of the attachment of C5b, C6, C7, C8, and C9 protein units. Especially, C9 protein forms a channel called Membrane Attack Complex (MAC), which can contribute to the formation of pore on the surface of bacteria. In this cycle, C3b is treated as the main controlling unit, which can be an important factor in this research.

2. Mannose-binding lectin

Mannose-binding lectins (MBL) are literally lectins; proteins, which form bonds with pathogens such as bacteria and parasites. Once they attach on the surface carbohydrates of those foreign organisms, MBL and mannose-binding lectin-associated serine protease (MASP) trigger a chain reaction of producing complements, which also eventually produce membrane attack complex (MAC). Our body cells show distinguishable carbohydrates in comparison to that of general pathogens, preventing self-destruction by our own MBL[3].

B. Proteins related to Mannose-binding Lectin

1. Mannose-6-phosphate isomerase

Mannose-6-phosphate isomerase is a enzyme which catalyzes the mutual converting reaction between

F6P(Fructose-6-phosphate) and M6P(Mannose-6-phosphate)[4]. This reaction can contribute to the ordinary ceullular transport and distinguish the cell membrane, whether the organism is prokaryotic or eukaryotic[5]. It consists of 440 Amino Acid resources, and it helps the reaction between C2 and C3, which is the part of alternative pathway of mannose. It is also an important signal substance, which is important for lisosomes and endosomes, and it can cause a serious defection to lisosomes in the case of the lack of this protein[6][7].

2. Triosephosphate isomerase

Triosephosphate isomerase catalyzes the mutual reaction between dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate. This protein is significant in the efficient energy production. It is found in most organisms, including eukaryotes and some bacteria. It catalyzes important metabolic reaction in an extremely faster speed, compared to the natural state. Except for the diffusion of the substrate, this protein has been considered as one of the most perfect catalysts in most of organisms. When the lack of this protein is occurred, it can lead to the severe neurological disease called triosephosphate isomerase deficiency, which is the result of hemolytic antenna and other mutations.

Comparing these two proteins that have direct relevance with mannose, we assumed that it is reasonable to analyze the possibility for forming chemical bond between human MBL and mannose of *Helicobacter pylori*.

C. Helicobacter pylori

As it can be known from the name, it refers to the spiral-shaped bacteria that live in pylori. *Helicobacter pylori* makes urease enzyme that can disassemble the elements in the gastric mucus into ammonia and carbon dioxide. And *Helicobacter pylori* use this ammonia in neutralizing gastric acid to live in the stomach that bacteria can't live because of strongly acid. Australian scientists Barry Marshall and Robin Warren identified *Helicobater pylori* in 1982[8], who discovered that it was existed in a person with chronic gastritis and gastritis ulcers. Moreover, It has been found that it is related to the growth of duodenal ulcers and stomach cancer. *Helicobacter pylori* is the only pathogen among bacteria that has been revealed, that may cause malignant tumors[9].

D. Data Mining Methods

1. Decision Tree Algorithm

Decision tree algorithm is a type of decision support tool that schematized decision making rules and the consequences follow, using tree-like graph or model of decision. Decision tree is primarily aimed for finding a strategy that brings similar result with the anticipated goal in decision analysis. In tree-like graph, leaves present the class labels. In the other hand, branches signify the logical multiply between class labels and related features. Decision tree as a visualized and apparently presented method is often used for decision making procedures and decided decision. An algorithm that constructs decision tree primarily uses top-down method. Furthermore, variable values that divide a given data by most suitable criteria are selected in each steps. Since decision tree is easy to understand and interpret, works well in a broad range of data, but does not necessarily processed, it is suitable for comparative study of amino acids and DNA sequences in bioinformatics, especially in the our interested field.

2. Apriori Algorithm

Apriori algorithm is a general algorithm for extracting frequency measurements from given data sets. It is generally utilized in bioinformatics to decide predominant component in specific organisms; usually amino acids and DNA sequences. The frequency level shows the degree of influence among various elements. Apriori algorithm puts data sets in certain length of candidate sets; from length k to length k-1. These candidate data sets include data in configured length k, making this algorithm to choose the most frequent data in the candidate data sets[10]. Final results are expressed with frequency measurements. Results from apriori algorithm indicate correlation between data sets based on their frequency level. If two data sets classified under a specific situation showed the highest frequency level, it is possible to consider this results as an existence of correlation between those data sets. In bioinformatics, apriori algorithm gives not only the possibility of correlation but patterns in data sets. That is, pattern analysis based on apriori algorithm is utilized for detecting the similarities of biological data such as DNA sequences.

3. Support Vector Machine

Support Vector Machines are one of supervised learning models in machine learning that generally used for classification and regression analysis[11]. When a set of data that belongs to one of two categories are given, an SVM algorithm builds a non-probabilistic binary linear classifier that assigns new data into one category or the other based on a given data. The classified models are representation of boundary that separate data into two categories. In these models, SVM algorithm is an algorithm that seeks for a suitable boundary which divides categories by a gap with a greatest width. SVM algorithm can be both used in linear classification and non-linear classification. Since SVM algorithm can be utilized in non-linear classification, it classifies more accurately and sensitively. The ultimate aim for SVM algorithm is in deciding where a novel data belongs to under the assumption in which a given data individually existed in two classes. SVM, with the strength that mentioned above, usefully applied in problems taking place in real world. Furthermore, the model exceedingly decreases previously required amount of learning studies in dividing text and hypertext. In addition, the model plays a pivotal role in classifying images. An experimental consequence released recently describes the fact that the model comes up with more accuracy in searching compared to Quarry Quantitative structure. Specifically, SVM presents the advantage when it comes to medical field that mainly classifies protein from already divided chemical substances up to approximately 90%. This gives an explanation why the model is not only primarily

applied to analysis of amino acids and DNA sequences in bioinformatics but also used in our interested field.

E. Datasets

In this research for comparing each proteins of *Helicobacter pylori* and Homo sapiens; mannose-6-phosphate isomerase and triosephosphate isomerase, amino acids sequences derived from NCBI data were used. ALM80176.1 and NP_002426.1 are each *Helicobacter pylori* and Homo sapiens mannose-6-phosphate isomerase. EIE30003.1 and CAA49379.1 are each amino acid sequences of *Helicobacter pylori* and Homo sapiens triosephosphate isomerase.

III. RESULTS

A. Decision Tree Algorithm

Name	Rules	Frequency
Helicobacter pylori	pos5 = N	0.900
	pos4 = E	0.824
	pos1 = G	0.857
	pos4 = E	0.833
	pos2 = N	0.889
	pos4 = E	0.824
Homo sapiens	pos4 = T	0.857
	pos2 = L pos4 = L	0.800
	pos4 = C	0.800
	pos4 = T	0.800
	pos4 = T	0.833
	pos2 = R	0.800
	pos1 = E pos4 = F	0.800

Table 1. Mannose-6-phosphate isomerase Rule Extraction Under 5 window

For TABLE 1, we extracted rules which indicates frequency higher than 0.824 for *Helicobacter pylori* and 0.800 for Homo sapiens, comparing noticeable rules between two organisms. According to the rules of *Helicobacter pylori* mannose-6-phosphate isomerase, which were extracted under 5 window, glutamic acid (E) at position 4 were extracted in high frequency, and is regarded as an important factor for differentiating *Helicobacter pylori* from Homo sapiens, since Decision Tree Algorithm were used to extract rules which shows difference in comparison to other datasets. Also from rules of Homo sapiens, we may assume that threonine (T) at position 4 were considered as a distinguishing factor.

For TABLE 2, 3, 4, and 5, we extracted rules which indicates frequency higher than 0.800.

Name	Rules	Frequency
Helicobacter pylori	pos1 = Y	0.833
	pos1 = S	0.800
	pos1 = S	0.900
	pos1 = Y	0.800
	pos1 = I	0.857
	pos1 = T	0.800
	pos1 = S	0.818
	pos3 = R	0.833
	pos3 = N	0.800
Homo sapiens	pos1 = P	0.875
	pos1 = Q	0.800
	pos1 = P	0.857
	pos5 = R	0.800
	pos1 = P	0.857
	pos1 = A	0.800
	pos3 = D	0.833

Table 2. Mannose-6-phosphate isomerase Rule Extraction Under 7 window

According to TABLE 2, tyrosine (Y) at position 1 from *Helicobacter pylori* is observed as a factor which may have worked as a standard of comparison for each organisms. Also Proline (P) at position 1 from Homo sapiens could be regarded to be a factor which indicates its distinguishable characteristic in amino acids sequences.

Name	Rules	Frequency
Helicobacter pylori	pos6 = G	0.800
	pos9 = K	0.800
	pos9 = F	0.800
	pos2 = F	0.800
	pos9 = L	0.875
Homo sapiens	pos6 = V	0.833
	pos7 = V	0.800
	pos2 = A	0.875
	pos9 = A	0.833
	pos9 = R	0.800
	pos9 = V	0.800

Table 3. Mannose-6-phosphate isomerase Rule Extraction Under 9 window

According to TABLE 3, various rules were extracted while there are no common amino acid sequence between two organisms. It is considered that *Helicobacter pylori* and Homo sapiens may have distinguishable characteristics in mannose-6-phosphate isomerase.

Name	Rules	Frequency
Helicobacter pylori	pos2 = K	0.800
	pos2 = K	0.818
	pos2 = K	0.875
	pos1 = Y	0.800
	pos2 = I	0.800
Homo sapiens	pos2 = C	0.833
	pos2 = A	0.813
	pos2 = C	0.800
	pos2 = A	0.800
	pos2 = V	0.800
	pos2 = D	0.800
	pos1 = E	0.833
	pos1 = T	0.833
	pos1 = H	0.800

Table 4. Triosephosphate isomerase Rule Extraction Under 5 window

According to TABLE 4, ricin (K) at position 2 from *Helicobacter pylori* could be considered as a standard of comparison for each organism. Also, cysteine (C) at position 2 and alanine (A) at position 2 from Homo sapiens are possible factors for differentiating standards between two organisms. In addition, since the existence for difference in the number of extracted rules; 5 for *Helicobacter pylori* and 9 for Homo sapiens, it is considered that the two group of amino acids sequences may have distinguishable features.

However, experiment for triosephosphate isomerase rule extraction under 7 window was resulted as a failure in extracting rules.

Name	Rules	Frequency
Helicobacter pylori	pos2 = F	0.800
	pos9 = K	0.800
	pos9 = F	0.800
	pos5 = A	0.800
	pos5 = I	0.800
	pos9 = L	0.875
	pos9 = E	0.800
Homo sapiens	pos2 = A	0.818
	pos5 = G	0.833
	pos7 = V	0.833
	pos6 = V	0.833
	pos9 = P	0.800
	pos9 = V	0.800
	pos9 = A	0.857

Table 5. Triosephosphate isomerase Rule Extraction Under 9 window

According to TABLE 5, no rules were shown as common factors for differentiating two organisms. It could be regarded to indicating various characteristics existing in each amino acids sequences from two datasets.

B. Apriori Algorithm

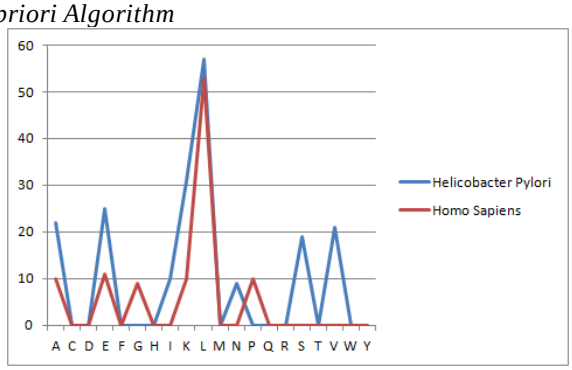

Figure 1. Mannose-6-phosphate isomerase frequency Under 5 window

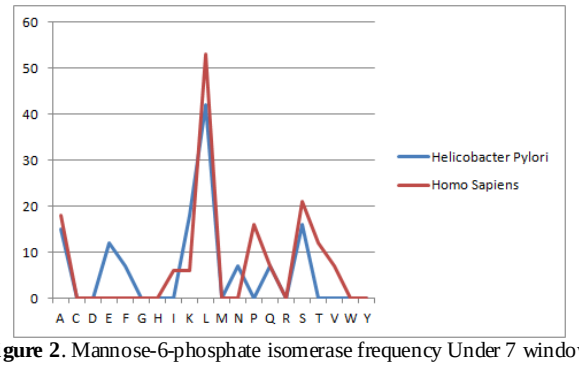

Figure 2. Mannose-6-phosphate isomerase frequency Under 7 window

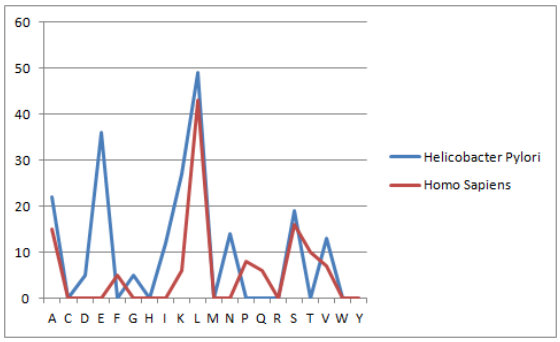

Figure 3. Mannose-6-phosphate isomerase frequency Under 9 window

Figure 1, 2 and 3 shows the frequency of mannose-6-phosphate in Helicobacter Pylori and Homo Sapiens. In the case of Helicobacter Pylori, the frequency of E, K, L, S type amino acid was showed to be higher than the others. Except the existence difference of V type amino acid, the data from three windows showed a similar tendency of amino acid distribution.

In the case of Homo Sapiens, the frequency of L, P type amino acid was showed to be higher than the others. This data showed a definite similarities compared to Helicobacter Pylori, which is the high frequency of L type amino acid. Also, the data from window 5 showed a big difference related to the existence of S and V type amino acid.

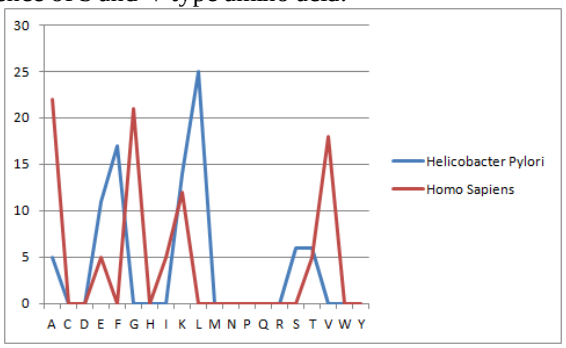

Figure 4. Triosephosphate isomerase frequency Under 5 window

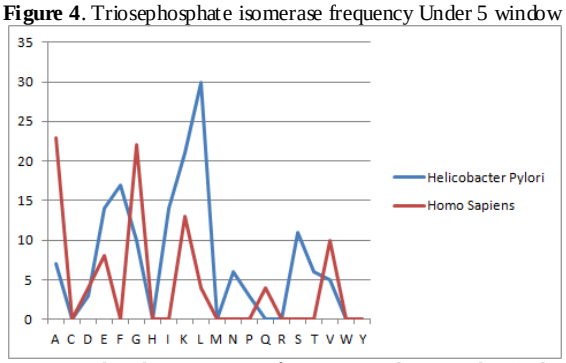

Figure 5. Triosephosphate isomerase frequency Under 7 window(Helicobater Pylori)

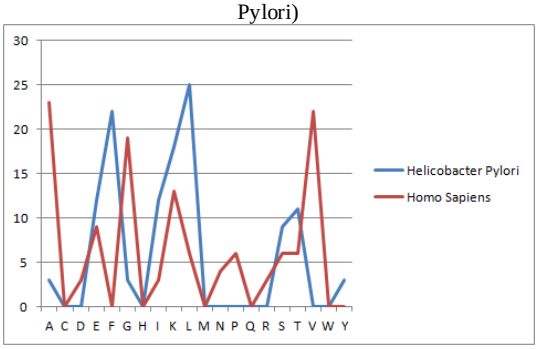

Figure 6. Triosephosphate isomerase frequency Under 9 window

The three above graphs state the frequency of Triosephospahte isomerase. In the case of Helicobacter Pylori, F, L, S, T type amino acid show relatively high frequency than other types of amino acid. However, different data under separate windows showed the difference of existence of I, M, P, V type amino acid.

In the case of Homo Sapiens, A, E, G, K, V type amino acid showed the relatively high frequency, compared to the others.

Also, various types of amino acid such as I, Q, R, S, T were shown in some windows, but not in all ones.

Also, some special cases of tendency has been observed in the following chart, which has the potential to be analyzed using Decision Tree. These data are not considered on the above graphs.

Window	Amino Acid Frequency Result
7 Window	amino1=E amino7=K 3 ==> cleavage=yes 3
	amino1=S amino7=I 3 ==> cleavage=yes 3
	amino2=E amino4=L 3 ==> cleavage=yes 3
	amino4=L amino5=E 3 ==> cleavage=yes 3
	amino5=E amino7=I 3 ==> cleavage=yes 3
9 Window	amino1=I cleavage=yes 3 ==> amino7=E 3
	amino1=I amino7=E 3 ==> cleavage=yes 3
	amino1=I 3 ==> amino7=E cleavage=yes 3
	amino8=F cleavage=yes 3 ==> amino2=F 3
	amino2=F cleavage=yes 3 ==> amino8=F 3
	amino2=F amino8=F 3 ==> cleavage=yes 3
	amino8=F 3 ==> amino2=F cleavage=yes 3
	amino2=F 3 ==> amino8=F cleavage=yes 3
	amino3=S cleavage=yes 3 ==> amino7=E 3
	amino3=S amino7=E 3 ==> cleavage=yes 3
	amino3=S 3 ==> amino7=E cleavage=yes 3

Table 6. Unique tendency of Triosephosphate isomerase frequency(Helicobacter Pylori)

Moreover, similar to Helicobacter Pylori, triosephosphate isomerase of Homo Sapiens also showed some unique tendencies of amino acid results in some windows, which is stated in Table 7.

Window	Amino Acid Frequency Result
9 Window	amino2=I 3 ==> amino5=K 3
	amino2=I cleavage=yes 3 ==> amino5=K 3
	amino2=I amino5=K 3 ==> cleavage=yes 3
	amino2=I 3 ==> amino5=K cleavage=yes 3
	amino5=G amino6=E 3 ==> cleavage=yes 3

Table 7. Unique tendency of Triosephosphate isomerase frequency(Homo Sapiens)

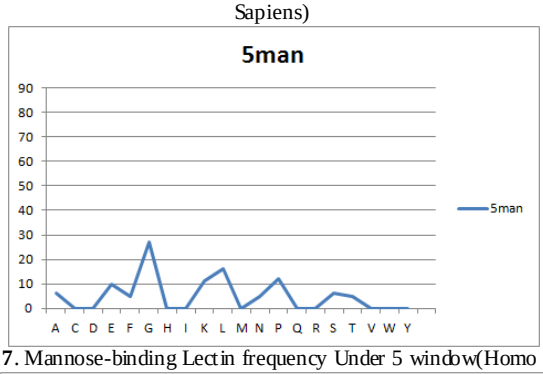

Figure 7. Mannose-binding Lectin frequency Under 5 window(Homo Sapiens)

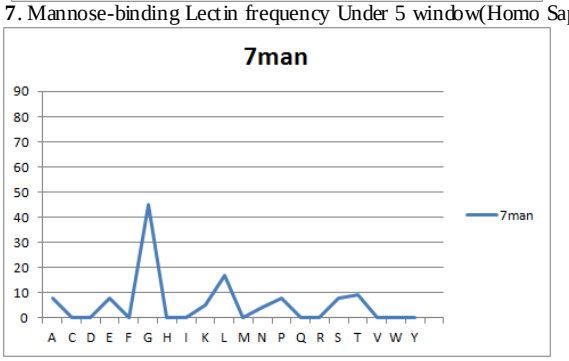

Figure 8. Mannose-binding Lectin frequency Under 7 window(Homo Sapiens)

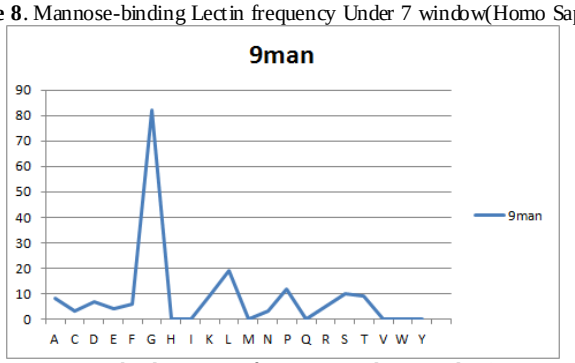

Figure 9. Mannose-binding Lectin frequency Under 9 window(Homo Sapiens)

Figure 7, 8, and 9 show the high frequency of G type amino acid. Also, all data under three windows showed the high rate of similarity compared to other types of amino acid.

C. SVM

It is considered to classify datasets of amino acid sequences of *Homo sapiens* and *Helicobacter pylori* in two group; linear classification using normal function and non-linear classification using kernel function. Testing each datasets 10 times for each functions, it is necessary to calculate the average. The average of accuracy values from these classification testing by each functions are shown in the Tables.

	Normal function	Polynomial function	R.B.F	sig
Helicobacter pylori	56	57	75	44.5
Homo sapiens	54.25	65.75	76.25	51.25

Table 7. The Accuracy Values of *Helicobacter pylori* and *Homo sapiens* for each function Under 5 Window

	Normal function	Polynomial function	R.B.F	sig
Helicobacter pylori	48	60.5	74	51.5
Homo sapiens	51.334	59.667	78.666	56

Table 8. The Accuracy Values of *Helicobacter pylori* and *Homo sapiens* for each function Under 7 Window

	Normal function	Polynomial function	R.B.F	sig
Helicobacter pylori	54	52	73	43
Homo sapiens	60.625	67.5	78.25	53.75

Table 9. The Accuracy Values of *Helicobacter pylori* and *Homo sapiens* for each function Under 9 Window

Similarity between *Homo sapiens* and *Helicobacter pylori*. Classification by SVM in the condition of window 9, which the result is more accurate than other window sizes(See TABLE 9), showed that the accuracy of *Helicobacter pylori* was higher than *Homo sapiens*. That means the datasets of *Homo sapiens* was better classified than *Helicobacter pylori* and that indicates that the similarity of *Helicobacter pylori* is much higher.

IV. CONCLUSION

Based on three different algorithms; Decision Tree Algorithm, Apriori Algorithm and Support Vector Machine, it is showed that the results were varied.

According to Decision Tree Algorithm, Helicobacter pylori and Homo sapiens showed distinguishable features of amino acids in mannose-6-phosphate isomerase and triosephosphate isomerase, since rules that are regarded as a key factor of differentiation were extracted from both two types of proteins. However, the absence of extracted rules for triosephosphate isomerase rule extraction under 7 window could be considered as existence of possible similarity between two organisms.

In addition, according to Apriori Algorithm, leucine (L) showed highest frequency in mannose-6-phosphate isomerase of both organism. However, in triosephosphate isomerase, two organisms indicated different characteristics; phenylalanine (F), leucine (L) were frequently showed from *Helicobacter pylori* while alanine (A), glycine (G), valine (V) were observed in high frequency from *Homo sapiens*. This result could be considered that two organisms have similar amino acid sequence for mannose-6-phosphate isomerase but different for triosephosphate isomerase.

Additionally, according to SVM, since RBF which generally shows nearly 100% of accuracy for classification showed about 75% of accuracy, it is considered that two organisms have many common factors.

In conclusion, we could not derive common explanation for the possibility of forming chemical bond between MBL and *Helicobacter pylori* mannose, since the tree algorithm showed different results as mentioned above. However, further research for examination of chemical bond formation through enzyme-substrate reaction could draw supporting conclusions. Also, the results of unique tendency of amino acid sequence frequency from Apriori Algorithm could be researched for factors we might have overlooked. Analysis on the possibility of chemical bond formation through comparing the similarities between two proteins which have direct relevance with mannose could suggest the potential of utilizing proteins of immune system such as MBL for targeting foreign factors. Researchers could improve the targeting treatments by considering immunological approach which includes application of other proteins to targeting therapy; represented as major histocompatibility complex (MHC), immune globulins, and cytokines.

REFERENCES

[1] Ji X, Gewurz H, Spear GT (February 2005). "Mannose binding lectin (MBL) and HIV". *Mol. Immunol.* **42** (2): 145–52
[2] Masihi KN (July 2001). "Fighting infection using immunomodulatory agents". Expert Opin Biol Ther. 1 (4): 641–53.
[3] Sompayrac, Lauren. *How the Immune System Works*. Malden, MA: Blackwell Pub., 2008.
[4]EBI Database, IPRO16305 Mannose-6-phosphate Isomerase
[5]Berg, Tymoczko, Stryer (2007). Biochemistry (6th ed.). New York, New York: W.H. Freeman &Co.
[6]Cleasby, Anne et al. (1996). "The X-ray crystal structure of phosphomannose isomerase from Candida albicans at 1.7 Å resolution"
[7]Jaeken, J., Matthijs, G. (2001). "Congenital Disorders of Glycosylation.". Annu. Rev. Genomics Hum. Genet. 2: 129–151.
[8]Marshall BJ (1983). "Unidentified curved bacillus on gastric epithelium in active chronic gastritis". 《Lancet》 **1** (8336): p.p.1273–1275.
[9]Blaser MJ (2006). "Who are we? Indigenous microbes and the ecology of human diseases"(PDF). *EMBO Reports*. **7** (10): 956–60. .
[10] 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 JBM* 03.06 (2015): 49-53.
[11]Cortes, C.Vapnik, V. (1995). "Support-vector networks". Machine Learning20 (3): 273–297.

Seo Jun Kim was born in Gangneung, South Korea in 1999. He is studying biology and computer science in Hankuk Academy of Foreign Studies, and interested in gene expression of immune system and influenza virus research. He published a paper about H5N1, H5N3, H7N9 hemagglutinin and neuraminidase sequence analysis using decision tree algorithm on International Conference on Chemical and Process Engineering.

Won Jong Lee was born in Daejeon, South Korea in 1999. He is now in Hankuk Academy of Foreign Studies. He is interested in immunology and bioinformatics using data mining. He is majoring in bioinformatics and dental biomaterials in the academy and conducted a presentation about observation of dissociation and reformation of hydroxyapatite depending on pH and surfactant on The Korean Society for Biomaterials.

Juho Jung was born in Seongnam, South Korea in 1998. He is studying computer science in Hankuk Academy of Foreign Studies. He is interested in bioinformatics using data mining and influenza virus research. He also majoring immunology in the academy and he is planning to research on artificial intelligence.

Taeseon Yoon was born in Seoul, Korea, in 1972. He received the 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, 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.