Assessment of Criteria Regarding CpG Islands of Their Role as Gene Marker

Haewon Kim, Jongjun Lee, Seonghui Yu, Junhyung Bae, Yeonho Jung, and Taeseon Yoon

***Abstract*—CpG islands are clusters of CG-rich DNA sequences, approximately half of which exist in 5' region of human housekeeping genes. Using various AI algorithms, previous studies classified CpG islands based on the criteria: length ≥ 200 bp, %GC ≥ 50%, and ObsCpG(Observed CpG)/ExpCpG(Expected CpG) ≥ 0.60. They showed that CpG islands overlap the promoter of all human housekeeping genes and over half of all tissue-specific genes.**

Using CpGIE(CpG Island Explorer), we evaluated suggested criteria for the search of CpG island of human gene chromosome 18, 19, 20. Manipulating three different criteria: length, %GC, and ObsCpG/ExpCpG rate, we found most appropriate criteria for CpG detection. One of the most important processes of our assessment is to decide whether the new criteria effectively exclude *Alu* repeats. Moreover, by assorting association type: promoter-relation, within-relation, end-relation, we confirmed that CpG islands defined by the new criteria showed better function as gene markers.

***Index Terms*—Bioinformatics, CpG island, gene marker.**

I. Introduction

In DNA sequence of human gene, CG appears less frequently than other dinucleotide sequences such as GG, GA, GT, and GC. However, in the 5' end of gene, the frequency rises up to the expected value. The clusters of the dinucleotide observed in this part of gene are called CpG islands. There are about 45,000 CpG islands in the human gene. These CpG islands are considered to be crucial in cell differentiation and regulation of gene expression. [1] As most CpG islands are located near the promoters, these CpG islands are employed as gene markers which indicate where the promoters and first exons of human genomes are. [2] Thus, plotting CpG islands in human gene can show the association between themselves and promoters, and, furthermore, the effect of CpG dinucleotide on human gene expression.

Criteria employed to identify CpG islands have been improved over time. In the primary studies, criteria for a CpG island were a DNA sequence longer than 200 bp; %GC ≥ 50%; CpGobs(observed)/exp(expected) ≥0.6 [3].

CpGobs(observed)/exp(expected) means the number of CpG divided by the number of C × the number of G × N. (N is the total number of nucleotides in the sequence which is analyzed) Recently, more strict criteria are developed. Takai and Jones proposed criteria of length longer than 200 bp; %GC ≥55%; CpGobs/exp≥0.65 [4] Despite concerns that new criteria could also exclude already identified CpG islands, reports showed that the proportion of genes with CpG islands was almost the same under the old and the generally accepted criteria [5].

Previous studies employed classification methods such as AI algorithms, neural networks. The classification became more sophisticated. The result showed an improvement from the previous studies. Also, the experiment was taken *on homo sapiens* chromosome 18, 19, 20.

II. Purpose

A. Expected Results and Purpose of Experiment

Our goal is to predict housekeeping genes associated with CpGisland effectively and accurately. We tried to make new criteria which can precisely predict the association between genes and CpG islands. With a computer program counting CpG islands on different parts of genes, we expected to find new criteria that can define promoter-associated CpG islands more effectively.

III. Means of Experiment

A. Procedure of Equipment

1) CpGIE (CpG island explorer)

The equipment used for our experiment is CpGIE(CpG Island Explorer). CpGIE is a java program developed by Yong Wang *et al.* [6] It is a tool for CpG island searching. It has interface panel in which criteria used for CpG island search can be manipulated. Sequences can be loaded on the program, and the program graphically demonstrates CpG islands on the genes. Compared with other programs, it showed better performance than other prorams such as CpGPlot, and CpGProD.[6]

2) Algoritm

To search for CpG island, we used CPGIE-a java program for CpG island analysis, developed by Takai and Jones [4]. The main processes to manipulate CPGIE are followed:

a) Edit the input data

b) Sorting the initial CpG islands out

- Set up the criteria, In this study, we searched CpG islands according to three different criteria: length ≥ 500bp, G+C content ≥ 60%, and CpG, o/e ratio ≥ 0.70 (newly suggested criteria 1) and length ≥ 500bp, G+C content ≥ 60%, and CpG, o/e ratio ≥ 0.60(newly suggested criteria 2) and length ≥ 500bp, G+C content ≥ 55%, and CpG, o/e ratio ≥ 0.65(generally accepted criteria by previous studies) .

Manuscript received November 10, 2013; revised January 10, 2014.

The authors are with Hankuk Academy of Foreign Studies, Wangsan-ri, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, Korea (e-mail: janny4615@naver.com, ljj960819@gmail.com, seonghui12@gmail. com, hjweebee1@naver.com, gwi03186@naver.com).

DOI: 10.7763/IJCEA.2014.V5.412

- Considering the DNA sequence which was inputted by the first input steps, minimize the length of window. Before identify the satisfying criteria, a moving window have to adjust at least (CpG o/e ratio) $*$ (minimum length) /16 CpG dinucleotides. This procedure will eliminate mathematical CpG islands which were caused by an enhanced component proportion ratio of G over C, or the transposed [4]
- The beginning points are marked if some regions in a window were verified as a CpG islands. Then, at intervals of 10 nt, the window moves ahead to find the first region in the window that is not a CpG island. It is also marked that the stop position of the CpG islands are in the last window.
- If there is a sequence that fulfill the criteria in the middle of marked area (between start site and stop site), it is classified as a initial CpG island. Otherwise, until it meets satisfying criteria, both ends of the sequence will be cut at the intervals of 1 nt simultaneously.
- The window keeps moving at interval of 1 nt. To reiterate the proposed processes (ii, iii, iv) enable CpGIE to search all the initial CpG islands in their location order.

c) Integrate the initial CpG islands

The recorded initial CpG islands are mostly superimposed due to the subtle moving steps of previous steps. Draw a comparison of the end site of an initial CpG island and the start site of the circumjacent backward CpG island, and assemble the superimposing and close-spaced (<100 nt in distance) initial CpG islands in a group. To create a final CpG island, initial CpG islands in the equal group will be employed.

The location that the CpG islands possess in the DNA sequence-between the start position and stop position of the initial CpG island- is decided by the criteria. If there are cutting processes, it will keep performing to find the final CpG islands in the region.

Although the cutting process of the sequence in the region was performed, no final CpG island is discovered in some cases. In these respects, the initial CpG islands usually in the middle of the region show the final island in this group.

3) Datasets

We excerpted all the information about the contigs in human chromosomes from NCBI (http://www.ncbi.nlm.nih.gov). The contigs in the chromosome 18, 19, 20 we used in this study were: NT_010966.14, NT_010895.14, NT_011109.16, NT_077812.2, NT011362.10, NT_011333.6, NT_025215.4. Information about the genes and contigs with their start and end sites, transcription orientation and evidence code used in this study was acquired from the NCBI.

4) Connection between CpG Island and Genes

Using CpGIE, we searched CpG islands in chromosomes 18, 19, and 20 based on the generally accepted criteria (by previous studies): length ≥ 500bp, %GC content ≥ 55%, and CpG, o/e ratio ≥ 0.65. And repeat the process on the basis of newly suggested criteria: length ≥ 500bp, G+C content ≥ 60%, and CpG, o/e ratio ≥ 0.70(new criteria 1) and length ≥ 500bp, G+C content ≥ 60%, and CpG, o/e ratio ≥ 0.60(new criteria 2). Similar to the genes in the contigs, the CpG islands appeared depending on their start site and end site location.

According to suggested criteria and generally accepted criteria for the search of CpG islands, we marks the genes, CpG islands on the contigs. We categorized the connections between CpG islands and genes into promoter relation, within relation and end relation regarding its association.

To visualize the correlation between CpG island location and gene location, we made a graph with sequence NT_025215.4 in chromosome 20. We use Microsoft Excel program to array nucleotide sequences, and marking them. In Microsoft Excel program, base pairs are arrayed 70 pairs per row, so we make chart to display quotient and remainder of start and end point when they are divided by 70. Also, we set 70 base pairs as one window, and transfer windows with 70 pairs each. We got CG percentage, marked gene and marked CpG islands by each window. Then we could get the graph by marking CpG islands and Gene induced through each other criteria.

IV. Results and Discussion

A. New Criteria Are More Effective in Finding CpG Islands Related to Genes

The role of CpG islands as gene markers are assessed by their relativity to genes. There are simply two parts in chromosomes. One of them is gene section, and the other one is the part that remains after excluding gene section from the chromosome. By using CpGIE as mentioned before, the total number of CpG islands in the whole contig and the number of CpG islands that is related to genes were checked.

This observation was progressed based on two newly suggested criteria which are criteria 1: length ≥ 500bp, G+C content ≥ 60%, and CpG, o/e ratio ≥ 0.60, criteria 2: length ≥ 500bp, G+C content ≥ 60%, and CpG, o/e ratio ≥ 0.70 and one generally accepted criteria : length ≥ 500bp, G+C content ≥ 55%, and CpG, o/e ratio ≥ 0.65 , and this will be referred to as established criteria. [6] The CpG islands related to genes include promoter-related, within-related, and end-related CpG islands (see Table I-IV).

TABLE I: (A) Figures in the Brackets Are the Numbers of Total CpG Islands in the Contigs. Figures Outside the Brackets Are the Numbers of CpG Islands that are Related to Genes

	New criteria 1	New criteria 2	Established criteria
NT_010859.14	96 (158)	93 (120)	131(178)
NT_010966.14	130 (171)	129 (143)	167(220)
NT_011109.16	995 (1160)	759 (945)	1067(1329)
NT_077812.2	70 (72)	54 (56)	73(78)
NT_011333.6	161 (167)	104 (106)	163(176)
NT_011362.10	173 (454)	134 (339)	180(536)
NT_025215.4	1 (7)	0 (5)	1(7)

TABLE I: (B) GENE-RELATION RATE OF THREE DIFFERENT CRITERIA

	New criteria 1	New criteria 2	Established criteria
NT_010859.14	0.620	0.775	0.736
NT_010966.14	0.760	0.902	0.759
NT_011109.16	0.858	0.803	0.803
NT_077812.2	0.972	0.964	0.936
NT_011333.6	0.964	0.981	0.926
NT_011362.10	0.381	0.395	0.336
NT_025215.4	0.143	0.000	0.143

TABLE II: THE NUMBER OF PROMOTER-RELATED, WITHIN-RELATED, END-RELATED CPG ISLANDS, DETERMINED BY NEW CRITERIA 1

Chromosome	Contig	Promoter	Within	End	Promoter Relation Rate
18	NT_010859.14	56	39	1	58.333
18	NT_010966.14	70	60	0	53.846
19	NT_011109.16	310	635	50	31.156
19	NT_077812.2	24	42	4	34.286
20	NT_011333.6	41	119	1	25.466
20	NT_011362.10	31	68	2	30.693
20	NT_025215.4	0	1	0	0
	* Criteria are %GC 60%, O/E 0.60, length 500 bp				

TABLE III: THE NUMBER OF PROMOTER-RELATED, WITHIN-RELATED, END-RELATED CPG ISLANDS, DETERMINED BY NEW CRITERIA 2

Chromosome	Contig	Promoter	Within	End	Promoter Relation Rate
18	NT_010859.14	54	38	1	58.065
18	NT_010966.14	71	57	1	55.039
19	NT_011109.16	240	497	22	31.621
19	NT_077812.2	23	29	2	42.593
20	NT_011333.6	35	68	1	33.654
20	NT_011362.10	71	62	1	52.985
20	NT_025215.4	0	0	0	0
	* Criteria are %GC 60%, O/E 0.70, length 500 bp				

TABLE IV: THE NUMBER OF PROMOTER-RELATED, WITHIN-RELATED, END-RELATED CPG ISLANDS, DETERMINED BY ESTABLISHED CRITERIA

Chromosome	Contig	Promoter	Within	End	Promoter Relation Rate
18	NT_010859.14	63	66	2	48.092
18	NT_010966.14	85	81	1	50.898
19	NT_011109.16	379	654	34	35.520
19	NT_077812.2	26	41	6	35.616
20	NT_011333.6	39	120	4	23.926
20	NT_011362.10	87	90	3	49.333
20	NT_025215.4	1	0	0	100.000
	* Criteria are %GC 55%,O/E 0.65, length 500bp				

The role of CpG islands as gene markers is especially highlighted because CpG islands are highly related to the promoter section of genes [5]. In other words, the rate of promoter-related CpG islands is outstanding, compared to other types of gene-related CpG islands. The number of different types of gene-related CpG islands was counted.

Even if the number of newly defined CpG islands dropped, it is shown that new criteria are more efficient gene markers, assessed by two ways. First, gene relation rate ((The number of gene-related CpG islands)/(The number of total CpG islands in the whole contigs)) arose dramatically in two new criteria. Since one common difference of two new criteria with established criteria is G+C content, which is more stringent than the established criteria, it is expected that higher G+C content contributes to more efficient gene marking by using CpG islands.

Second, promoter relation rate ((The number of promoter-related CpG islands)/(The number of total CpG islands in the whole contigs)) arose dramatically in new criteria. It was especially higher in new criteria 2, as it was showed in NT_011109.16 and NT_011255.14, NT_077812.2. This tendency can be contributed to two factors. First, within-related and end-related CpG islands might have lower G+C content than promoter-related CpG islands, as promoter-related CpG islands better satisfied both new criteria, whose only common difference is the rise in G+C content criteria. Also, new criteria 2 might show better promoter relation rate, owing to higher o/e rate. According to previous studies, higher G+C content and o/e ratio are effective in filtering *Alu* repeats [4]. Thus, more stringent G+C content and o/e ratio criteria might contribute to new criteria becoming better criteria of CpG islands as gene markers.

Thus, gene relation rate and promoter relation rate proves that two new criteria are better tools of detecting gene-promoter section by using CpG islands.

B. Comparision of New Criteria 1 and New Criteria 2 Regarding CpG Islands' Role as Gene Markers

Comparing new criteria with established criteria using Gene relation rate and Promoter relation rate, we conclude that new criteria 2 shows better effectiveness than new criteria 1 in predicting CpG islands which function as gene markers. Higher gene relation rate and promoter relation rate tells that new criteria 2 are better criteria. New criteria 2 have higher o/e rate. Higher o/e rate has an effect on criteria's function as gene marker. It is especially well shown from higher promoter relation rate. This noticeably improved effectiveness is thought to be a result of higher o/e rate, which is reported to exclude *Alu* repeats well [4], [5]. Because *Alu* repeats are often miscounted as CpG islands that are associated with genes, this advance in excluding *Alu* repeats shows clear difference between two criteria as their role as gene markers (see Table V-VII and Fig. 1- Fig. 4).

TABLE V: ESTABLISHED CRITERIA

Start	End	GC%	O/E(CpG)	Length	start int	start mod	end int	end mod
96587	97316	61.6	0.65	730	1379	57	1390	16
98900	99761	62.6	0.65	862	1412	60	1425	11
104520	105067	55	0.68	548	1493	10	1500	67
105571	106079	55.1	0.65	509	1508	11	1515	29
115183	115734	63.8	0.65	552	1645	33	1653	24
131357	132441	61.6	0.66	1085	1876	37	1892	1
191388	192814	61.8	0.69	1427	2734	8	2754	34

TABLE VI: NEW CRITERIA 1

Start	End	GC%	O/E(CpG)	Length	start int	start mod	end int	end mod
96493	97332	60.8	0.61	840	1378	33	1390	32
98724	99943	61.4	0.65	1220	1410	24	1427	53
104713	105251	60	0.6	539	1495	63	1503	41
106752	107264	73	0.6	513	1525	2	1532	24
115201	115869	61.8	0.6	669	1645	51	1655	19
131144	132440	61.5	0.62	1297	1873	34	1892	0
191558	192913	61.8	0.69	1356	2736	38	2755	63

TABLE VII: NEW CRITERIA 2

Start	End	GC%	O/E(CpG)	Length	start int	start mod	end int	end mod
96671	97305	61.5	0.7	635	1381	1	1390	5
98909	99698	63.5	0.7	790	1412	69	1424	18
115084	115593	60	0.71	510	1644	4	1651	23
131468	132381	61.4	0.7	914	1878	8	1891	11
191558	192555	62.5	0.71	998	2736	38	2750	55

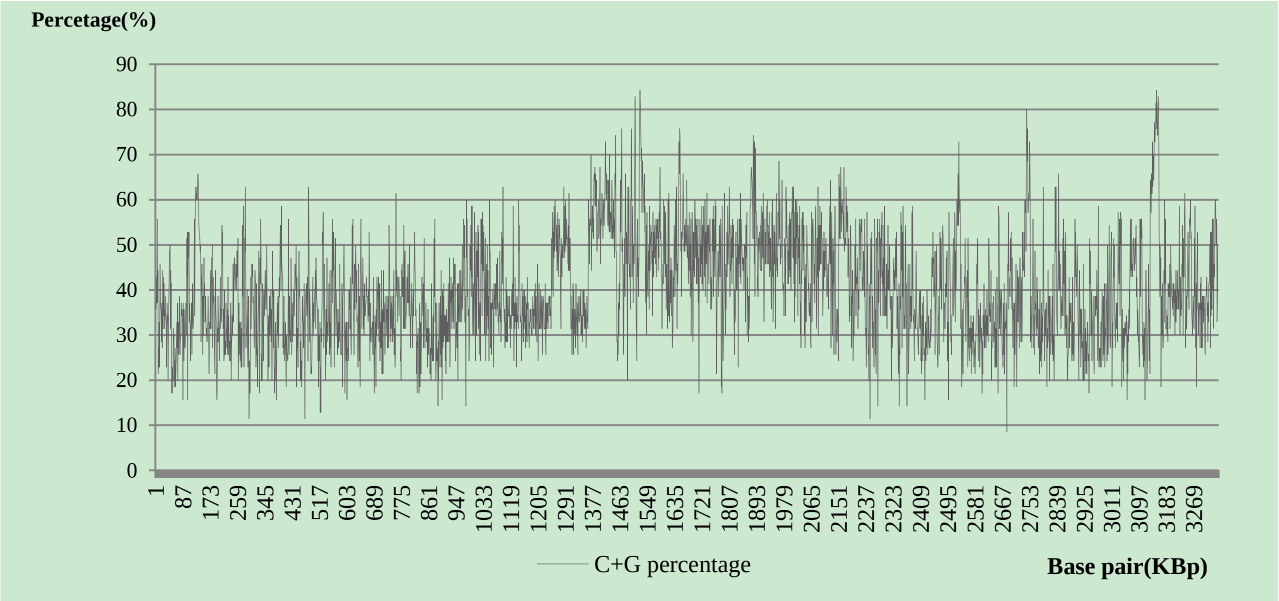


Fig. 1. CG percentage of NT_025215.4.

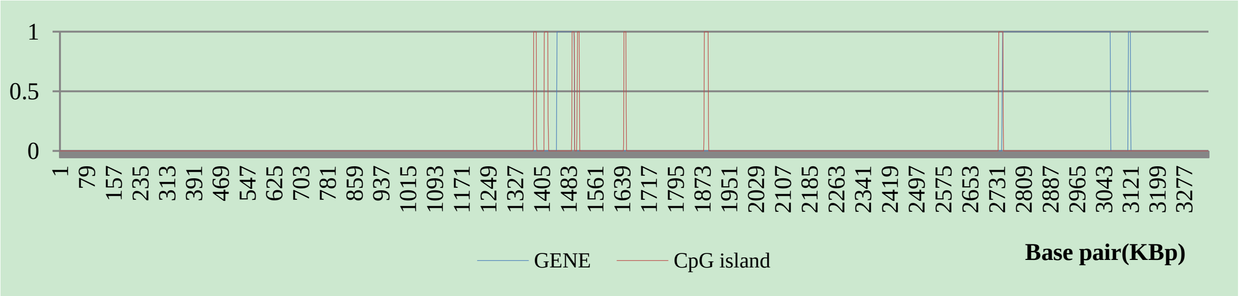


Fig. 2. Established criteria.

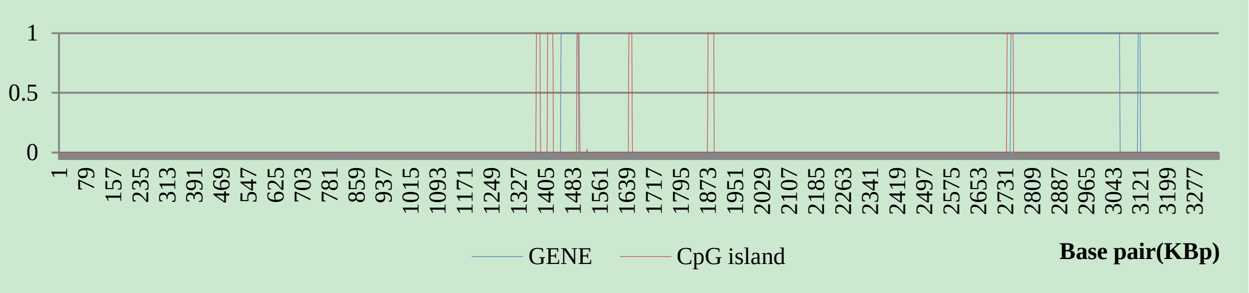


Fig. 3. New criteria 1.

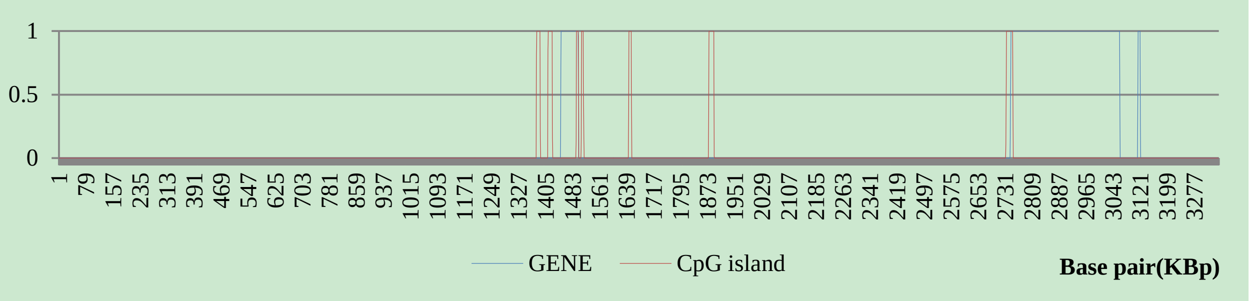


Fig. 4. New criteria 2.

V. Conclusion

Our study have proved two newly suggested criteria's possibility of being promoted in future studies, as both criteria have shown significant ability of finding gene-marker CpG islands. Higher G+C content was proved to improve gene-relation rate, compared to established criteria. Moreover, by comparing our two newly suggested criteria, we found out that the higher o/e ratio is, the higher promoter-relation rate their CpG islands show. This might be explained by exclusion of Alu repeats, and further research seems to be needed.

We could reach to conclusion that gene location and CpG island position recorded in some tendency. If CG portion is higher than the others, the higher possibility of CpG islands positioning, and gene positioning either. Also, CpG island location and Gene location usually overlapped, as we can assure the tendency to be positive.

Acknowledgment

Special thanks to Taeseon Yoon, who invited all of us into the world of bioinformatics.

References

[1] A. P. Bird, "DNA methylation patterns and epigenetic memory," *Genes Dev.*, vol. 16, pp. 6–21, 2002.

[2] M. Scherf, A. Klingenhoff, and T. Werner, "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach," *J. Mol. Biol.*, vol. 297, pp. 599–606, 2000.

[3] M. Gardiner-Garden and M. Frommer, "CpG islands in vertebrate genomes," *J. Mol. Biol.*, vol. 196, pp. 261–282, 1987.

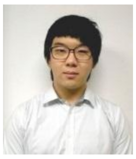
[4] D. Takai and P. A. Jones, "Comprehensive analysis of CpG islands in human chromosomes 21 and 22," *Proc. Natl Acad. Sci.* , USA, vol. 99, pp. 3740–3745, 2002.

[5] F. Larsen, G. Gundersen, R. Lopez, and H. Prydz, "CpG islands as gene markers in the human genome," *Genomics*, vol. 13, pp. 1095–1107, 1992.

[6] Y. Wang and F. C. C. Leung, "An evaluation of new criteria for CpG islands in the human genome as gene markers," *Bioinformatics*, vol. 20, pp. 1170-1177, 2004.

[7] F. Antequera and A. Bird, "Number ofCpGislands and genes in human and mouse," *Proc. Natl Acad. Sci., USA*, vol. 90, pp. 11995–11999, 1993.

Seonghui Yu was born in 1996. She is currently a student in science major of hankuk academy of foreign studies. She is interested in bio-science and gene analysis associated with clinical psychology which detect human behavior pattern.



Yeonho Jung was born in 1996. He is currently a student in science major of hankuk academy of foreign studies. He is currently interested in bio-science and modeling process to find most efficient way of analyzing and solving given problems.

Junhyung Bae was born in 1996. He is currently a student in science major of hankuk academy of foreign studies. He is interested in bio-science and applying bio-infomatics to analyzing gene and finding new facts about human genes.

Jongjun Lee was born in 1996. He is currently a student in science major of hankuk academy of foreign studies. He is mostly interested in applying bioinformatics and gene information to basic medical research.

Haewon Kim was born in 1996. She is currently a student in science major of Hankuk Academy of Foreign Studies. She is interested in bio-science and analyzing genetic facts or phenomenon related to humans