http://dx.doi.org/10.17661/jkiiect.2021.14.6.487

Requirements Analysis and System Design for the Implementation of the Gut Microbiome Analysis Platform

Wiseman Lim*, Sanghyuk Ma***, Sangbae Ma**, Hyoungmin Choi***

장내미생물 분석 플랫폼 구현을 위한 요구사항 분석 및 시스템 설계

임복출*, 마상혁***, 마상배**, 최형민***

Abstract The analysis method of the microbiome has been evolving for a very long time, and the industrial field has grown rapidly with the start of human genome analysis 20 years ago. As continuous research continues, related industries have grown together, and among them, Illumina of the US has been leading the popularization of DNA analysis by developing innovative equipment and analysis methods since its establishment in 1998. In this paper, 'AiB Index', 'AiB Chart' using statistical process control and log-scale technique to analyze the gut microbiome analysis methodology and implement an algorithm that can analyze minute changes in the minor strains that can be overlooked in the existing analysis methods. want to implement. From the data analysis point of view, we proposed a platform for analyzing gut microbes that can collect fecal data, match and process gut microbes, and store and visualize the results.

Key Words : Analysis Platform, Linear/Log scale, Microbiome, Normal Distribution, Variance

1. Introduction

Ecosystem is a collective term for interacting organisms and the surrounding inanimate environment that interacts with and affects each other. When groups of organisms living in the same place and dependent on each other form a completely independent system, it can be called an 'ecosystem'. This means that interdependence and completeness are essential elements to form an ecosystem[1].

Humans also maintain a symbiotic relationship from inside and outside the human body to eukaryotes such as archaea, bacteria, and virus, as well as yeast and mold. The number of microorganisms in the human body is more than 10 times that of the total number of human cells, and the number of genes possessed by microorganisms is also several hundred times that of the human genome. It has been known that these microorganisms perform major functions through interaction with the human body, such as influencing the absorption and metabolism of nutrients in the human body, the maturation and development of the immune system and nervous system, and the occurrence and prevention of various diseases. The aggregate of all microorganisms naturally present in the human body is called the human microbiome[2].

*Strategy Research Divsion(CEO), Wise Data Labs Inc.
**Corresponding Author : AiBiotics, (sbma@aibiotics.kr)
***AiBiotics Inc.

Received December 16, 2021 Revised December 16, 2021 Accepted December 23, 2021

Therefore, in this paper, various methods for data analysis through collection, storage and pre-processing of gut microorganisms are studied, and the requirements of the analysis platform are analyzed and the system is designed. In order to verify the requirements, we intend to implement the minimum function and proceed with the evaluation.

2. Related Research

2.1 Microbiome

The human gut microbiome has a diverse community structure for each individual according to heredity, diet, and lifestyle from birth, and the gene aggregate of these microbiome is defined as the gut microbiome [2].

Microbiome is a compound word created by combining microbiota and genome. It can be said to be a microbial community that contains the entire genetic information. Full-scale research on the microbiome started in 2010, and interest grew as the human microbiome treatment was selected as one of the 'World's Top 10 Promising Future Technologies' at the World Economic Forum(Davos Forum) in 2014.

The fact that the microbiome has a great effect on the human body brought about a change in perception with the publication of a study by Dr. Jeffrey Gordon(University of Washington, USA) in 2006. Dr. Jeffrey Gordon, after injecting the feces of obese and lean mice into sterile mice, confirmed that the sterile mice injected with the feces of obese mice became obese faster than the sterile mice injected with the feces of lean mice. The results of a study that showed that the gut microbiome of skinny people and those of skinny people are different was published in Nature in 2006[3].

2.2 Gut Microbiome Analysis Methodology

Since 1970, more than 400-500 bacteria have been isolated from the intestine through anaerobic culture, but it has the disadvantage that more than 80% of the gut microorganisms cannot be cultured[4]. In 1977, Woese and Fox discovered a 16S rRNA molecule that reflects bacterial phylogenetic characteristics by coexisting conserved and variable sequences. Over the next 30 years, 16S rRNA has been playing a key role in understanding the structure of the gut microbial community[5].

For analysis based on 16S rRNA, QIIME and mother pipeline are mainly used for cluster comparison. The analysis process generates OTUs(operational taxonomic units) based on sequence groups by noise filtering, sorting, and classification of large sequences. Perform network analysis and diversity analysis by comparing with the following sample information. On the other hand, whole gene sequencing analyzes the function of the microbial community by comparing it with the existing genetic information database through noise filtering, assembly, and gene prediction[2].

For the study of the gut microbiome, DNA is directly extracted from human samples such as feces without isolating or culturing microorganisms. DNA of all microorganisms is mixed in the extracted DNA sample, and the entire gene is analyzed directly at the gene level by confirming the nucleotide sequence using the next-generation sequencing method. Through this approach, the entire microbial community present in the sample can be identified, and the metabolic processes and functions of the symbiotic microorganisms can be revealed in the community[6].

2.3 Statistical Process Control(SPC) & Linear Scale

SPC is a quality control method that monitors and controls processes using statistical methods[7]. SPC provides useful problem -solving tools to achieve process stability and improve functionality through reduced variability. Key tools include histograms, check sheets, Pareto charts, causal relationships, flowcharts, scatter plots, and control charts[8].

For example, in the semiconductor process field, especially when measuring and managing impurities in a semiconductor high vacuum chamber, statistical process control is being used. The change in the concentration of minute impurities in the vacuum chamber has a great effect on the very dense semiconductor process, but because of the dominant concentration of the major gas, the change in the concentration of the minute gas does not distinguish between normal and abnormal on the linear scale[9], but the concentration change measured value is converted to a log scale[9][10], and the change in the concentration of fine impurities in the normal and abnormal state is clearly measured. Since the proportion of major microbes accounts for more than 80% of the composition of gut microbes, it is difficult to measure the distribution and changes of microscopic microbes on a linear scale.

3. Requirements Analysis

3.1 Sample Group and Control Group

In this paper, Noble Bio's NBgene-Gut (NGBG-2, NBG-3, NBG-4, etc.) was used to collect intestinal microbes. Stool samples from patients and general users in hospitals and clinics were collected, and the collection method is as shown in [Fig. 1] below[11].

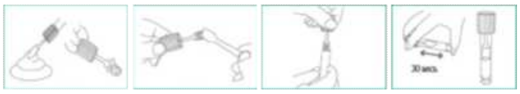

Fig. 1. Optimized microbiome collection solutions
Step 1. Collect feces with a sampling spoon
2. Push the collected feces on the top of the spoon as shown in the picture using the provided spatler. (Amount collected 1g)
3. Carefully insert the tube into the tube and grasp the cap.
4. Using the sus ball in the tube, shake the sample as shown in the figure to mix it well to make the sample uniform.

Microbiome Database refers to a database that analyzes the nucleotide sequence of many Bacteria existing in nature through several previous studies[Table 1]. It is the most common database of NCBI with many research results, and in this paper, the NCBI library used in the America GUT Project was used to secure the universality of the control group.

Table 1. Apply to Microbiome Database

Database	Category	Version	Update	Taxonomy
3BIGS	All	v1	2018.11	399,966
Silva	All	132	2018.04	425,098
Greengene	All(Qiime)	13.8	2014.04	203,452
eHOMD	Oral	15.11	2018.08	998
RDP	All	11.5	2016.09	Unknown
NCBI	All		now	29,461,785
ezTAXON	Chunlap		2018.01	63,240
Zymo	Unknown		2017.04	Unknown
PATRIC	Pathogen		2016.08	109,392

The collected samples were obtained by matching the sample data to the NCBI Micro biome Database using the MiSeq equipment [12] of Illumina, a representative NGS company, to obtain the results shown in [Table 2] below. In line one, 55BE020123, 55BR010011, etc. mean the ID of the sample and indicate the distribution of the amount by species.

Table 2. Result data using MiSeq

Species	55BE020123	55BR010011	...
Prevotella_copri	1676	24997	
Faecalibacterium_prausnitzii	4053	5838	
Bacteroides_dorei	21642	6138	
Streptococcus_salivarius	16	379	
Raoultella_ornithinolytica	4	4	
Prevotellamassilia_timonensis	63851	498	
omitted below			

Because the distribution of gut microbes accounts for more than 80% of the major strains, the extracted result data is used to measure minute changes in the minor strains, and the OTUs value is divided by the total number of data, and this is recalculated as a ratio. The results are shown in [Table 3] below.

Table 3. Sample data converted to ratio(%)

Species	55BE020123	55BR010011	...
Prevotella_copri	0.7%	22.8%	
Faecalibacterium_prausnitzii	1.8%	5.3%	
Bacteroides_dorei	9.7%	5.6%	
Streptococcus_salivarius	0.0%	0.3%	
Raoultella_ornithinolytica	0.0%	0.0%	
Prevotellamassilia_timonensis	28.5%	0.5%	
omitted below			

Phylum level, the largest classification standard, is currently divided into 12 divisions, and Bacteriodetes, Firmicutes, and Actinobacteria are most commonly distributed in the Phyllum Level. In the detailed classification below the Phylum level, it is divided into sub-species of Class, Order, Family, Genus, and Species.

3.2 Abnormal Analysis of Microbiome Distribution using Statistical Process Control and Log Scale

A representative method to analyze the microbiome is to analyze diversity, and there are Observed OUT, ACE, Shannon, Simson, InvSimson, and Fisher as a method of expressing diversity.

In particular, the Shannon Index is designed to reflect both types and amounts of interpretation of biodiversity. In this paper, AiB Chart and AiB Index are defined based on the theorem of Shannon Index. Shannon Index takes the log value of the probability distribution of individual species and expresses it as an inverse number. Since the probability distribution is always a value less than 1, and taking a log of this value is expressed as a negative number, it is calculated using the log function of the probability value of each species, converted to an absolute value, and then summed. The formula is as follows.

H = -sum p_i log(b) p_i

- p_i : abundance of species i
- b : log base(2 or natural log)

The AiB Chart started with the assumption that 100 million (100,000,000) bacteria are present in 1g of sample by applying the

Shannon Index applied to the general ecosystem to the Microbiome. Since the total distribution of bacteria included in 1g is 1, multiply the probability value of each species by 100 million times to estimate the number of bacteria in the original sample. However, since the decomposition capacity of the NGS equipment is at the level of 10^5 per sample, taking this into account, the value obtained by multiplying the probability distribution of each species by 10^6 is calculated. The number of each species multiplied by one million times is assumed to be the number of each species included in 1g, and the sum of all values obtained by taking the log function of the number of each species becomes the final AiB Diversity Index. The formula is as follows.

$$\boxed{\text{AiB Index} = \sum_{i=1}^{n} Log(Pi \times 10^6)}$$

The AiB Chart is a chart expressing individual values of the AiB Index as a bar graph, as shown in [Fig. 2] below. The x-axis is arranged in the order of abundance ratio by species, and the y-axis is the Pi x 10^6 value expressed on a log scale.

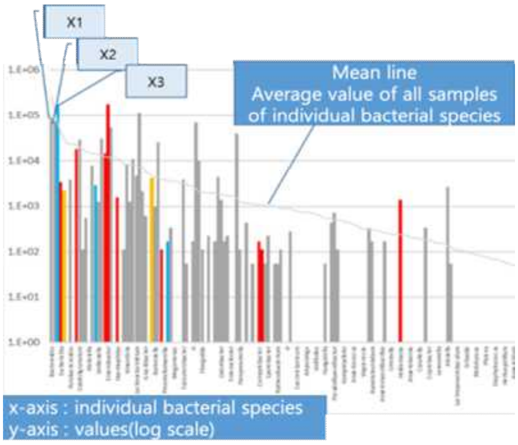

Fig. 2. Example of AiB Chart

The AiB Index is the sum of the area of individual bar graphs in the chart. That is, it means an area represented by a histogram. In the graph, one division on the y-axis is multiplied by 10, so compared to Bacteriodetes or Prevotella, which are generally most widely distributed, the bacterial species in the top 10% or less are almost 1/100. Since it is detected in a minute amount of about 1/1000, the discriminative power is inferior in a graph of a general scale. On the other hand, in the graph of log scale, the deviation of detailed species can be easily expressed. The [Fig 3] below is a graph when the above example is expressed in linear scale.

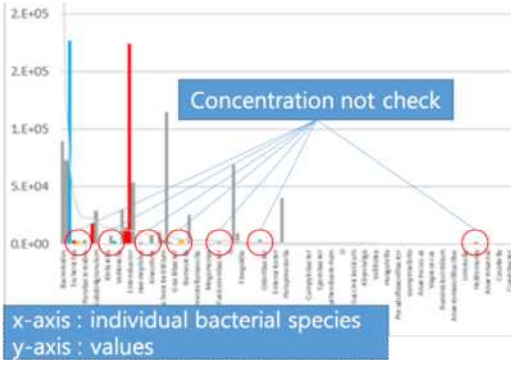

Fig. 3. Linier scale for comparison with AiB Chart

As a result, the AiB Chart was designed to reflect the characteristics of the microbiome with a high share of the major species, so that even a small amount of the species could be well expressed.

3.3 Automated Accuracy Improvement Algorithm based on Accumulation of Clinical Data

The association of microbiome with health has been studied through numerous experiments over the past 20 years, but it is very difficult to express the association in terms of absolute numbers of benefits and harms piecemeal.

However, the degree to which the distribution of beneficial bacteria is greater or less than the average is expressed as individual values. Bacteria that affect the digestive system, strains that affect immunity, and strains related to diabetes overlap and sometimes show opposite trends, and various experimental papers show various results.

For the AiB Index, standard deviation was used as a method to consider the size and directionality of the effect of each strain, and a Balanced Score Card applied with FMEA(Failure Mode Effect Analysis) was applied.

It is very dangerous and difficult to ensure consistency to simply judge how beneficial and harmful bacteria are by concentration. In particular, it is impossible to diagnose a disease simply by the concentration of harmful bacteria. However, the role of the microbiome was reflected as a criterion for judging normal and abnormal by looking at how significantly the combined concentrations of bacteria affecting specific diseases differ from normal[average].

A methodology for indexing the concentration deviation, that is, the weight, of the microbiome defined for each area was presented. In the AiB Chart, the deviation between the position of each bar graph and the average value means statistically deviating. As a method of judging normal and abnormal using the distribution of gut microbes, which cannot be determined, attention was paid to excess in the center of deviation. Considering that disease-causing bacteria are more affected by 'excess' than 'deficiency', attention was paid to the excess of bacteria of unknown function.

According to the accumulation of microbiome analysis data, the opinions of clinicians are collected on the relationship between bacteria and disease, and the weights such as excess or correlation are reflected in the system, and the automated scoring table is updated, which adds new weights to the analysis system. It is configured to operate with values. The [Table 4] below is an example of a scoring table.

Table 4. Scoring Table for concentration deviations in the microbiome(example)

Phylum	Genus	Species	IBD	DM
Firmicutes	Lactobacillus	Casei	0.1	0.3
Firmicutes	Anaerobium	acetethylicum	0	1
Firmicutes	Saccharofermentans	acetigenes	0	0
Firmicutes	Sporanaerobacter	acetigenes	0.1	0.4
Actinobacteria	Bifidobacterium	Infantis	0	-0.3
Actinobacteria	Bifidobacterium	adolescentis	0	0
Actinobacteria	Collinsella	aerofaciens	0.1	-0.1
Bacteroidetes	Bacteroides	acidifaciens	0.1	0.3
omitted below				

4. Prototyping and Evaluation

4.1 Design of Analytsis Platform

The analysis platform should be web-based and should be easily accessible to users, and should provide functions to manage user convenience and authority. Comparing the past and present for the implementation method of the web platform, it can be defined as shown in [Fig. 4] below.

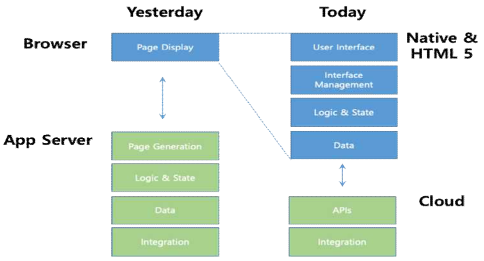

Fig. 4. Yesterday vs. Today of Web Platform

Recently, the service is provided in the form of SPA(Single Page Application), and by separating the screen part(F/E, Front End) and the data processing part(B/E, Back End), the rendering of the screen is minimized and data update is performed. The trend is to provide it so that it can be provided immediately according to the user's request.

For this purpose, various web frameworks exist, and the platform proposed in this paper uses the Vue.js(vuetify, vuechart, etc.) framework. Vue.js is a screen development web framework oriented to the MVVM(Model, View, ViewModel) pattern as shown in [Fig. 5] among UI screen development methods. It is a simple, easy-to-write, and component-based framework.

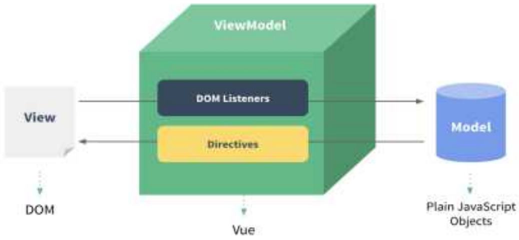

Fig. 5. Vue.js' MVVM Pattern(velog.io blog content)

In addition, it is a framework with the strengths of React and Angular, which are representative web frameworks, and is used in the construction of various web platforms in Korea.

The proposed platform is designed as shown in [Fig. 6] below by using the latest technologies and web frameworks of the web platform to configure the screen unit, data processing unit, and database as follows.

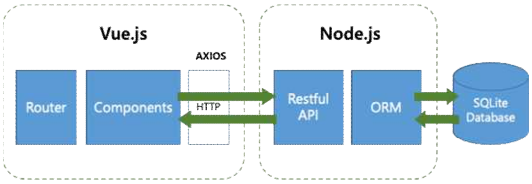

Fig. 6. Microbiome Analysis Platform Architecture

4.2 Prototyping of Analysis Platform

The roadmap was composed of three phases of advancement in consideration of the phased implementation of the analysis platform, development period, and cost.

As shown in [Fig. 7] below, in the first phase, an Open API for data interworking from analysis equipment, a screen for visualizing analysis results such as user information management and diversity analysis, and a backend processor for data processing are developed.

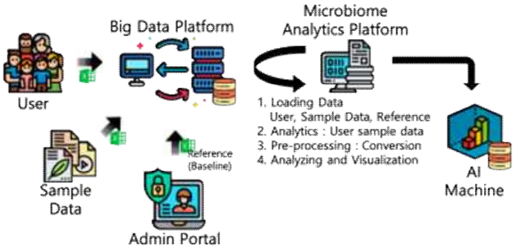
Big Data Platform
Microbiome
Analytics Platform
User
1. Loading Data
User, Sample Data, Reference
2. Analytics : User sample data
3. Pre-processing : Conversion
4. Analyzing and Visualization
Reference
(Baseline)
Sample
Data
AI
Machine
Admin Portal

Fig. 7. 1st phase Microbiome Analysis Platform

In this paper, some of the functions of the first phase of the analysis platform were implemented. Fecal sample data was collected from NBgene-Gut, and analysis data of the collected samples were extracted through MiSeq of Illumina. Based on the extracted data, mapping and conversion with reference data to provide users with the AiB charts, etc., and statistical data extraction and visualization were performed.

To implement the analysis platform, Ubuntu was used as the server system, and the latest version of Node.js was installed. Bio-fe for screen development and bio-stat projects for data processing were created and managed. The commands below show how to install and run each project.

```
(1)Frontend Project
- Vuetify add
npm install
npm install -g @vue/cli
npm install vue-modal-window
vue add vuetify

vue-cli-service serve (or) npm start
```

```
(2)Backend Project
npm install

node bin/www
```

4.3 Evaluation and Verification

The performance of the microbiome analysis platform can be how meaningful data is derived, correlation with disease, and influence with gut microbiome. In the field of analysis of the popular Gut Microbiome, the performance evaluation of diversity indicators is the key.

However, the analysis platform proposed in this paper reflects the requirements of the Gut Microbiome analysis platform as a result of prototyping, and how useful the system provided through this is for users. For the evaluation of the analysis platform for usefulness, it was evaluated based on several derived empirical cases.

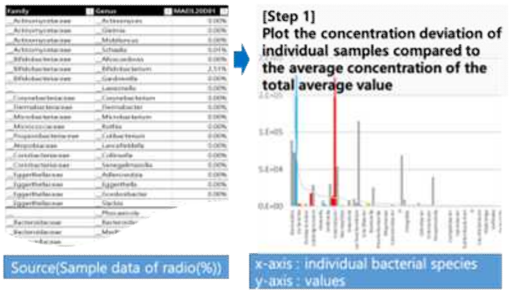

Fig. 8. Visualization for collected gut microbiome#1

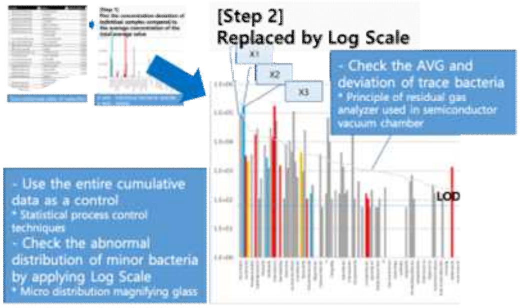

Fig. 9. Visualization for collected gut microbiome#2

The above [Fig. 8], [Fig. 9] is an application example of the visualization technology applied with semiconductor process management technology and Shannon Index, and even the smallest strains can be analyzed by applying the limit line and log scale using the mean and standard deviation.

The second is an example of system use as an 'abnormal' detection tool using SPC techniques as shown in [Fig. 10] below. The normal distribution of individual concentration data(individual histogram) for each sample of about 300 species of Genus level is expressed in one graph. Comprehensive analysis is possible by defining data with deviations(bias) from the normal distribution as ‘abnormal’. In the graph, the red dotted line indicates deviation from the normal distribution for the species exceeding the average line, which means the blue individual histogram, and visualizes this to analyze the ‘abnormal’ species at a glance.

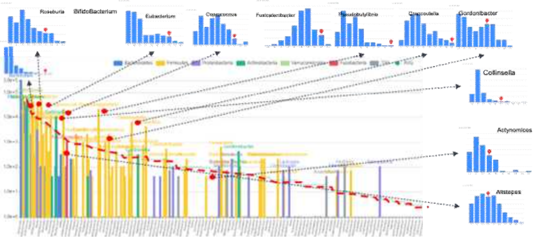

Fig. 10. Integrated Histogram Graph

5. Conclusion

The Gut Microbiome analysis platform proposed and implemented in this paper is a way to improve the problems of existing analysis platforms and methodologies, and it is a method to compare experimental group and control group, abnormal analysis through microbiome distribution, statistical process control technique, and log scale. The integrated normal distribution graph of the used Genus-level strains and the management of the scoring table according to the increase in clinical data were proposed. Currently, we have verified the requirements required by the Gut Microbiome analysis platform at the level of prototyping that implements some of the functions of the first phase.

For future research, we intend to implement the Gut Microbiome analysis platform planned in three phases, and study the software configuration and algorithm of the goals of each stage. In addition, the AiB Index for algorithm improvement and various analysis methodologies of the microbiome are fused and combined to define an index that has not been found in the correlation analysis with diseases and to study the relationship between the microbiome.

REFERENCES

[1] Fierer, Noah and Ferrenberg, Scott and Flores, Gilberto E. et al. From Animalcules to an Ecosystem: Application of Ecological Concepts to the Human Microbiome, Annual Review of Ecology, Evolution, and Systematics, 43(1):137-155, 43(1), 2012

[2] You, H. J., Lee, S., & Ko, G. Concepts and strategies of the human gut microbiome research, The Korean Journal of Public Health, 52(1), 2015

[3] Turnbaugh PJ. Ley RE. Mahowald MA. Magrini V. Mardis ER. Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 444:1027-1031, 2006

[4] Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. ProcNatl Acad Sci USA, 74:5088-5090, 1977

[5] Nocker A, Burr M, Camper AK. Genotypic microbial community profiling: a critical technical review. Microbial Ecol,

54:276-289, 2007

[6] Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes; a new frontier for natural peoducts. Chem Biol 5:R245-9, 1998

[7] Oakland, J. S. Statistical process control. Routledge. 2007

[8] Felix C. Veroya. Introduction to Statistical Process Control; A Problem Solving Process Approach. bookboon.com. 2014

[9] Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. Stat Med. Aug;7(8):889-94, 1998

[10] Cain, Clarence P., Noojin, Gary D., Manning, Lonnie. A Comparison Of Various Probit Methods For Analyzing Yes/No Data On A Log Scale. 1996

[11] http://noblebio.co.kr/

[12] https://sapac.illumina.com/systems/sequencing-platforms/miseq.html

저자약력

임 복 출(Wiseman Lim) [정회원]

- 2015년 : 중부대학교 일반대학원 정보과학과(공학박사)
- 2013년 : 중부대학교 인문산업대학원 정보과학과(공학석사)
- 1999년~2019년 : (주)비티비솔루션 스마트서비스 본부장 등
- 2019년~현재 : (주)위컴즈, (주)와이즈데이터랩 대표이사

〈관심분야〉
빅데이터컴퓨팅, 클라우드컴퓨팅, 빅데이터분석, 장내미생물 및 유전체 분석 등

마 상 혁(Sanghyuk Ma) [정회원]

- 2004년 : 고신대학교 대학원 소아과학 전공(의학박사)
- 2000년 : 고신대학교 대학원 소아과학 전공(의학석사)
- 1990년 : 경북대학교 의과대학
- 1995년~현재 : 창원 파티마병원 소아청소년과 주임과장
- 2020년~현재 : ㈜에이아이바이오틱스 대표이사

〈관심분야〉
헬스케어, 예방의학, 감염병/면역/비만 건강 빅데이터, 장내미생물 및 유전체 분석 등

마 상 배(Sangbae Ma) [정회원]

- 1993년 : 한양대학교 화학공학과(학사)
- 1993년~2003년 : 삼성전자 제조혁신그룹
- 2003년~2010년 : ㈜와이드칩스 외주관리/마케팅 부장
- 2010년~2016년 : ㈜실리콘화일 사업기획 본부장
- 2016년~2019년 : ㈜옵토레인 생산관리 본부장
- 2019년~현재 : ㈜에이아이바이오틱스 대표이사

〈관심분야〉
블록체인, 클라우드컴퓨팅, 빅데이터 분석, 장내미생물 및 유전체 분석 등

최 형 민(Hyoungmin Choi) [정회원]

- 2003년 : 호서대학교 정보통신공학과(학사)
- 2003년~2006년 : SK Hynix 품질보증팀(QA)
- 2006년~2015년 : ㈜실리콘화일 사업기획 파트장
- 2015년~2020년 : ㈜비트리 기획/마케팅 본부장
- 2020년~현재 : ㈜에이아이바이오틱스 영업마케팅 본부장

〈관심분야〉
장내미생물 및 유전체 분석, 반도체, 빅데이터 분석, 인공지능 등