Identification of Heavy Smokers through Their Intestinal Microbiota by Data Mining Analysis

Toshio KOBAYASHI[1,2]* and Kenji FUJIWARA[2,3]

[1] *Miyagi University, 2–2–1 Hatadate, Taihaku-ku, Sendai City, Miyagi 982-0215, Japan*
[2] *Riken, 2–1 Hirosawa, Wako, Saitama 351-0198, Japan*
[3] *Yokohama Rosai Hospital, Kozukue-cho, Kohoku-ku, Yokohama 222-0036, Japan*

Received September 27, 2012; Accepted November 24, 2012

The intestinal microbiota compositions of 92 Japanese men were identified following consumption of identical meals for 3 days, and collected feces were analyzed through terminal restriction fragment length polymorphism. The obtained operational taxonomic units and smoking habits of subjects were analyzed by a data mining software. The constructed decision tree was able to identify explicitly the groups of smokers and nonsmokers. In particular, 4 smokers, who smoked 20 cigarettes/day, i.e., heavy smokers, were gathered in the same group of the decision tree and were clearly identified. Related operational taxonomic unit were traced to understand the species of bacteria, but all were found to be uncultured bacteria.

Key words: human intestinal microbiota, restriction enzyme, operational taxonomic unit, smoking habit, heavy smoker, data mining analysis, decision tree

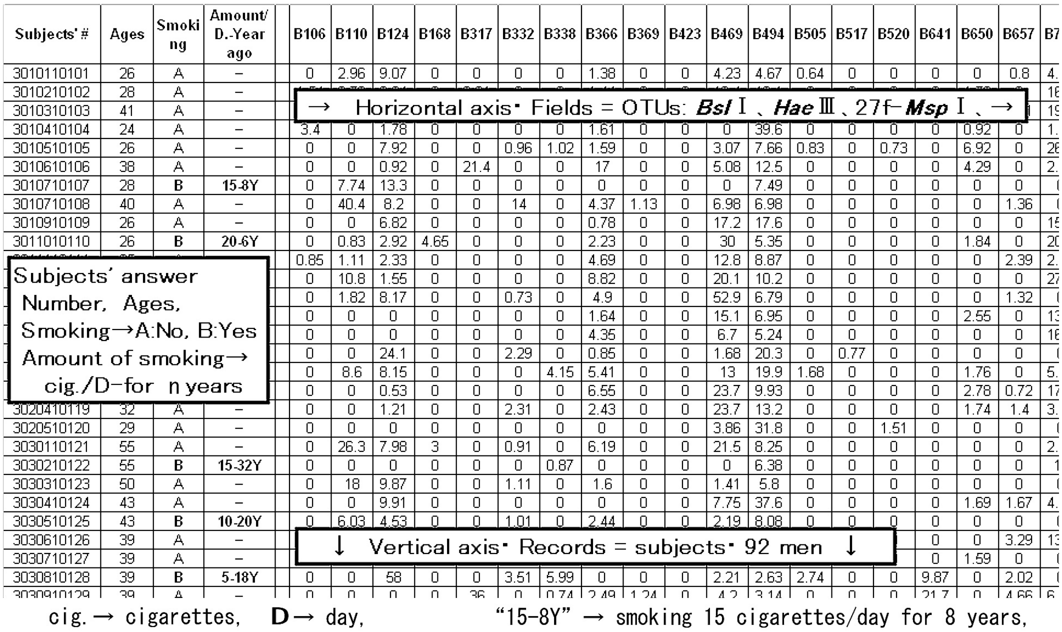
The human intestinal microbiota (HIM) is closely related to our health, and practical research on the relationship with the human immune systems and diseases is now being performed. Here we tried to apply data mining analysis (DM) to identify or discriminate the relation between the smoking habits of subjects and obtained HIM data from feces.

To avoid the influences of dietary factors, we designed identical meals (1,879 kcal/d), which were fed to 92 healthy male volunteers living in Japan for 3 days. All dietary components were controlled, and beverages were restricted to water, black coffee, or green tea in order to control carbohydrate intake. The ages and body mass indexes (BMI) of the subjects were 21–59 years (average: 36.8) and 17.3–30.1 kg/m^2 (average: 22.6), respectively. Fecal samples were analyzed by terminal restriction fragment length polymorphism (T-RFLP) using 3 primer restriction enzyme systems [1, 2].The reason for applying T-RFLP was as follows. First, the numerical data obtained from T-RFLP are reproducible, and second the processing is comparatively easy and reasonable for handling large numbers of subjects. Third, T-RFLP provides appropriate numbers of data for a subsequent numerical analysis, which requires a balance between the field number (horizontal axis) and records number (vertical axis). The studies were performed in accordance with the protocol approved by the RIKEN Research Ethics Committee, and the OTU data were accumulated by the Benno Laboratory, RIKEN.

Bacterial DNA was isolated from 40–100 mg of feces using the modified method described by Matsuki et al. [3]. Amplification of the fecal 16S rDNA, restriction enzyme digestion, size fractionation of the T-RFs and T-RFLP analysis were carried out as previously described [4–6]. PCR was performed with FAM-labeled 516f (5′-TGCCAGCAGCCGCGGTA-3′; *E. coli* positions 516-532) or 27f (5′-AGAGTTTGATCCTGGCTCAG-3′; *E. coli* positions 8-27) and the reverse primer 1510r (5′-GGTTACCTTGTTACGACTT-3′; E. coli positions 1510–1492). For the PCR products amplified with the 516f primer, the resulting 16S rDNA amplicons were further treated for 1 hr with 2 U of *Bsl*I or *Hae*III (New England Biolabs, Ipswich, MA, USA), and for those amplified with the 27f primer, the resulting 16S rDNA amplicons were treated for 1 hr with 1 U of *Msp*I (TaKaRa Bio Inc., Otsu, Shiga, Japan). The digestion products were fractionated using an automated sequence analyzer (ABI PRISM 3130xl DNA Sequencer, Applied Biosystems, Carlsbad, CA, USA) and analyzed with the GeneMapper software (Applied Biosystems).

The obtained data for operational taxonomic units (OTUs) were abbreviated as B--- (---: base pair number) for *Bsl*I, HA--- for *Hae*III and M--- for 27f*Msp*I. The amounts for each OTU represent the fluorescence intensity and then concentrations of each OUT group. These OTU data are reproducible and can be used for further numerical analyses. A total of 80 OTUs were combined with the answers of 92 subjects and analyzed by

*Corresponding author. Toshio Kobayashi. Fax: +81 3-717-7398. E-mail: toskoba@attglobal.net



Subjects' #	Ages	Smoking	Amount/ D.-Year ago		B106	B110	B124	B168	B317	B332	B338	B366	B369	B423	B469	B494	B505	B517	B520	B641	B650	B657
3010110101	26	A	–		0	2.96	9.07	0	0	0	0	1.38	0	0	4.23	4.67	0.64	0	0	0	0	0.8
3010210102	28	A	–																			
3010310103	41	A	–																			
3010410104	24	A	–		3.4	0	1.78	0	0	0	0	1.61	0	0	0	39.6	0	0	0	0	0.92	0
3010510105	26	A	–		0	0	7.92	0	0	0.96	1.02	1.59	0	0	3.07	7.66	0.83	0	0.73	0	6.92	0
3010610106	38	A	–		0	0	0.92	0	21.4	0	0	17	0	0	5.08	12.5	0	0	0	0	4.29	0
3010710107	28	B	15-8Y		0	7.74	13.3	0	0	0	0	0	0	0	0	7.49	0	0	0	0	0	0
3010710108	40	A	–		0	40.4	8.2	0	0	14	0	4.37	1.13	0	6.98	6.98	0	0	0	0	0	1.36
3010910109	26	A	–		0	0	6.82	0	0	0	0	0.78	0	0	17.2	17.6	0	0	0	0	0	0
3011010110	26	B	20-6Y		0	0.83	2.92	4.65	0	0	0	2.23	0	0	30	5.35	0	0	0	0	1.84	0
					0.85	1.11	2.33	0	0	0	0	4.69	0	0	12.8	8.87	0	0	0	0	0	2.39
					0	10.8	1.55	0	0	0	0	8.82	0	0	20.1	10.2	0	0	0	0	0	0
					0	1.82	8.17	0	0	0.73	0	4.9	0	0	52.9	6.79	0	0	0	0	0	1.32
					0	0	0	0	0	0	0	1.64	0	0	15.1	6.95	0	0	0	0	2.55	0
					0	0	0	0	0	0	0	4.35	0	0	6.7	5.24	0	0	0	0	0	0
					0	0	24.1	0	0	2.29	0	0.85	0	0	1.68	20.3	0	0.77	0	0	0	0
					0	8.6	8.15	0	0	0	4.15	5.41	0	0	13	19.9	1.68	0	0	0	1.76	0
					0	0	0.53	0	0	0	0	6.55	0	0	23.7	9.93	0	0	0	0	2.78	0.72
3020410119	32	A	–		0	0	1.21	0	0	2.31	0	2.43	0	0	23.7	13.2	0	0	0	0	1.74	1.4
3020510120	29	A	–		0	0	0	0	0	0	0	0	0	0	3.86	31.8	0	0	1.51	0	0	0
3030110121	55	A	–		0	26.3	7.98	3	0	0.91	0	6.19	0	0	21.5	8.25	0	0	0	0	0	0
3030210122	55	B	15-32Y		0	0	0	0	0	0	0.87	0	0	0	0	6.38	0	0	0	0	0	0
3030310123	50	A	–		0	18	9.87	0	0	1.11	0	1.6	0	0	1.41	5.8	0	0	0	0	0	0
3030410124	43	A	–		0	0	9.91	0	0	0	0	0	0	0	7.75	37.6	0	0	0	0	1.69	1.67
3030510125	43	B	10-20Y		0	6.03	4.53	0	0	1.01	0	2.44	0	0	2.19	8.08	0	0	0	0	0	0
3030610126	39	A	–																	0	0	3.29
3030710127	39	A	–																	0	1.59	0
3030810128	39	B	5-18Y		0	0	58	0	0	3.51	5.99	0	0	0	2.21	2.63	2.74	0	0	9.87	0	2.02
3030910129	39	A	–		0	0	0	0	36	0	0.74	2.49	1.24	0	4.2	3.14	0	0	0	21.7	0	4.66

→ Horizontal axis· Fields = OTUs: *BslⅠ*、*HaeⅢ*、27f-*MspⅠ*、→

Subjects' answer
Number, Ages,
Smoking→A:No, B:Yes
Amount of smoking→
cig./D-for n years

↓ Vertical axis· Records = subjects· 92 men ↓

cig.→ cigarettes, **D**→ day, "15-8Y" → smoking 15 cigarettes/day for 8 years,

Fig. 1. A part of the obtained 2 dimensional excel data

data mining (DM) software (IBM-SPSS Clementine-14). Due to the large scale of the data, only a portion of the 2-dimensional Excel data is shown in Fig. 1 as an example. The subjects contained 16 smokers, and their smoking habits were abbreviated as number of cigarettes/ day-number of years, e.g., "5-2Y" means smoking 5 cigarettes/day for 2 years.

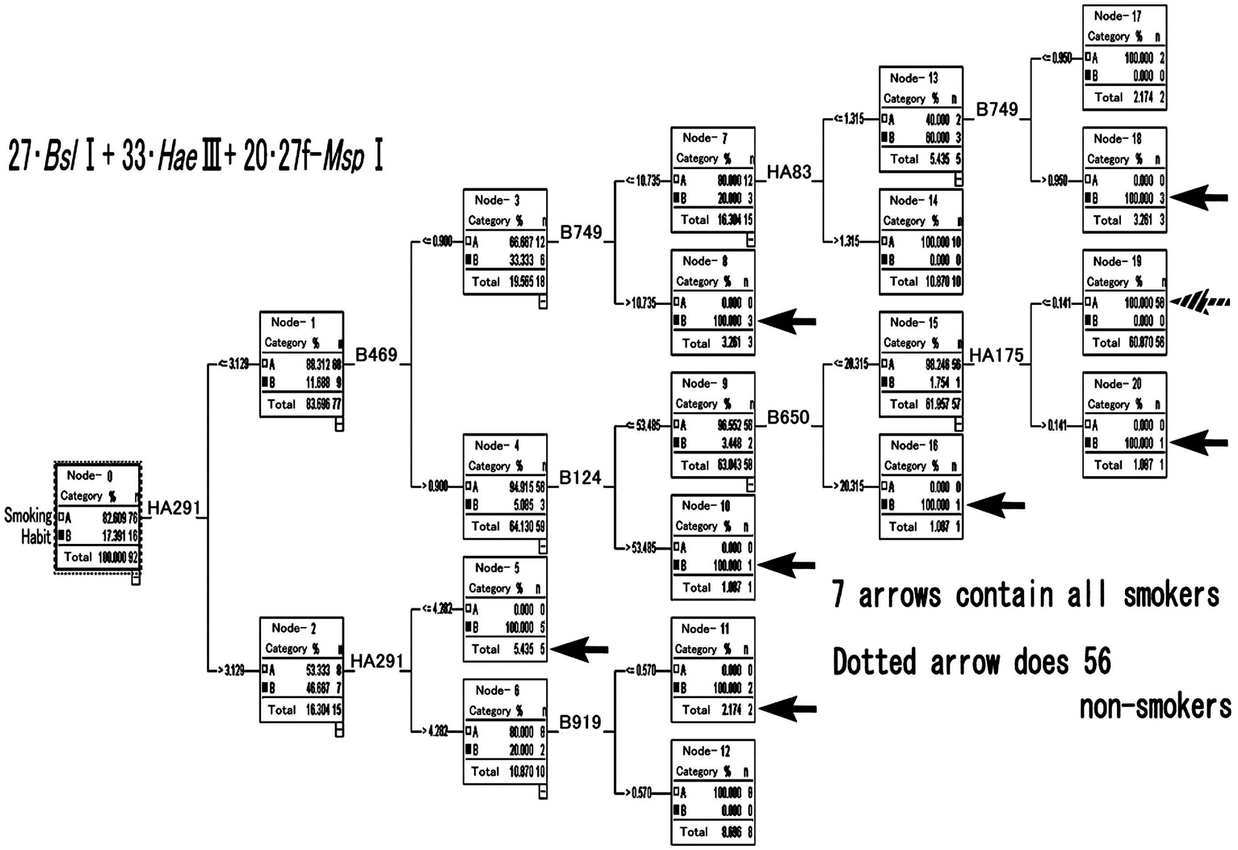
After the analyses, DM provided a decision tree[1] (Dt) as shown in Fig. 2, which identified explicitly the various groups of smokers. The left end of Fig. 2 is called the root node, i.e., the starting point of tree construction, and the Dt grew toward the right to divide the subjects. The details of the Dt and the pathway to reach the terminal node[2] indicated clearly the species and quantities of OTUs, which played a role in dividing the various smoking groups (i.e., node). We applied a dividing system using the Classification and Regression Tree (C&RT) approach, which is the most typical construction method for Dt, using the Gini coefficient between the smoking status and OTUs data and divided the records into two subsets so that the records within each subset were more homogeneous than in the previous subset. C&RT is quite flexible, and allows unequal misclassification costs to be considered compared with the other growing systems of DM. In Fig. 2, the 7 arrows indicated all the nodes of the 16 smokers (B: yes), and the dotted arrow indicated a node that contained 56 subjects, i.e., 74% of the nonsmokers (A: no) were gathered in the node. A major specialty of this method of DM is that it uses a single selected OUT for each step of Dt construction. In Fig. 2, only 8 OTUs out of 80 were utilized, with 2 OTUs, i.e., HA291 and B749, being applied twice, meaning that the other 72 were not used to construct the tree shape. Therefore, we can accept the fact that a large number of subjects were gathered in a node, i.e., 56 subjects in N-19. In other words, only these 8 OTUs were related in some way with the present smoking habits.

Paying attention to the nodes of smokers with arrows in Fig. 2, Table 1 showed detailed subject records for the 7 nodes, which included all 16 smokers, and compared the subjects' answers and DM predictions. All of the subjects who were habitual smokers were explicitly identified in the 7 nodes. In addition, in N-5 at the lower part of Table 1, all 4 heavy smokers, who smoked 20 cigarettes/day, e.g., 20-6Y, were gathered together. These facts actually indicated that the selected 8 OTUs were related in some way with the amounts of smoking and that some HIM were sensitive to the habits or characteristics of the individuals.

As for the pathway to reach N-5 in Fig. 2, utilization of an OTU, HA291 (*Hae*III-291), twice indicated a very close relation with heavy smokers, so we traced the species of bacteria. Simple tracing of HA291 with the

[1] decision tree: decision supporting pathway that makes use of a treelike graph

[2] terminal node: tree nodes that do not split further


27·*Bsl* I + 33·*Hae* III + 20·27f-*Msp* I
Smoking Habit
Node- 0 Category % n □A 82.609 76 ■B 17.391 16 Total 100.000 92
HA291
<=3.129
>3.129
Node- 1 Category % n □A 88.312 68 ■B 11.688 9 Total 83.696 77
B469
Node- 2 Category % n □A 53.333 8 ■B 46.667 7 Total 16.304 15
HA291
<=0.900
>0.900
Node- 3 Category % n □A 66.667 12 ■B 33.333 6 Total 19.565 18
B749
Node- 4 Category % n □A 94.915 56 ■B 5.085 3 Total 64.130 59
B124
<=4.282
>4.282
Node- 5 Category % n □A 0.000 0 ■B 100.000 5 Total 5.435 5
Node- 6 Category % n □A 80.000 8 ■B 20.000 2 Total 10.870 10
B919
<=10.735
>10.735
Node- 7 Category % n □A 80.000 12 ■B 20.000 3 Total 16.304 15
HA83
Node- 8 Category % n □A 0.000 0 ■B 100.000 3 Total 3.261 3
<=53.485
>53.485
Node- 9 Category % n □A 96.552 56 ■B 3.448 2 Total 63.043 58
B650
Node- 10 Category % n □A 0.000 0 ■B 100.000 1 Total 1.087 1
<=0.570
>0.570
Node- 11 Category % n □A 0.000 0 ■B 100.000 2 Total 2.174 2
Node- 12 Category % n □A 100.000 8 ■B 0.000 0 Total 8.696 8
<=1.315
>1.315
Node- 13 Category % n □A 40.000 2 ■B 60.000 3 Total 5.435 5
B749
Node- 14 Category % n □A 100.000 10 ■B 0.000 0 Total 10.870 10
<=20.315
>20.315
Node- 15 Category % n □A 98.246 56 ■B 1.754 1 Total 61.957 57
HA175
Node- 16 Category % n □A 0.000 0 ■B 100.000 1 Total 1.087 1
<=0.950
>0.950
Node- 17 Category % n □A 100.000 2 ■B 0.000 0 Total 2.174 2
Node- 18 Category % n □A 0.000 0 ■B 100.000 3 Total 3.261 3
<=0.141
>0.141
Node- 19 Category % n □A 100.000 56 ■B 0.000 0 Total 60.870 56
Node- 20 Category % n □A 0.000 0 ■B 100.000 1 Total 1.087 1

7 arrows contain all smokers

Dotted arrow does 56 non-smokers

Each □ is called "node". The left end node is called "Root-node", which is the starting point of tree construction. Dt was growing toward right side.

The marks, e.g. HA291, was the dividing OTU, of which numerical dividing point were shown.

Each node showed its component of subjects. Node-n was abbreviated as "N-n" in Table 1.

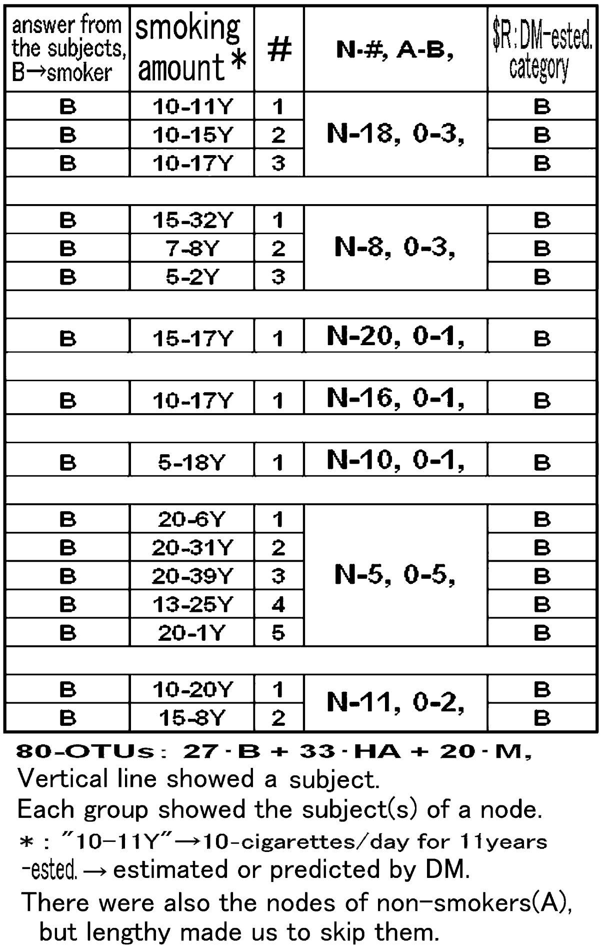
Fig. 2. Decision-tree (Dt) obtained by DM smoking habit with 80-OTUs

Microbial Community Analysis III web site and tools of Idaho University [7] revealed 1036 registered bacteria. Then by comparison of 3 other restriction enzymes, *Bsl*I, 27f-*Msp*I and 27f-*Alu*I with HA291, over 30,000 of bacteria species were scanned and crosschecked by accession number. Finally 28 bacteria were screened, but all were uncultured species, and some were identified as rumen bacteria and soil bacteria. However, this simplified the possibilities and indicated that not all 28 bacteria existed in the HIM of heavy smokers.

Looking closely at the lower right part of Fig. 2, we realized that there were 10 other subjects who had larger amounts of HA291. This meant that the heavy smokers in N-5 had higher intermediate amounts of HA291. The T-RFLP method contains various bacteria in an OTU, so HA291 was not a single bacterium. Then the lower right part of Fig. 2 showed some different species of bacteria.

Comparing our results with the former classification methods of HIM, the most unique point was the introduction and application of DM identification and predictive analyses. Previously, cluster analyses have been popularly applied for obtained OTUs [8, 9], but they suffered from the following 2 limitations. Namely, the first is that the cluster shows only some classified groups but did not show visible reasons for reaching the groups. Second, the obtained cluster is tightly attributable to the data, meaning that if a slight modification is made to add or subtract data, the next cluster will be very different from the previous one, i.e., each cluster lacks flexibility. On the other hand, according to the Dt, DM showed clear reasons for the tree construction, so sequential rolls of selected OTUs and simple utilization

Table 1. "Smoking Habit with 80-OTUs" comparison between the subject's answer and the DM-analyses



answer from the subjects, B→smoker	smoking amount *	#	N-#, A-B,	$R:DM-ested. category
B	10-11Y	1	N-18, 0-3,	B
B	10-15Y	2		B
B	10-17Y	3		B
B	15-32Y	1	N-8, 0-3,	B
B	7-8Y	2		B
B	5-2Y	3		B
B	15-17Y	1	N-20, 0-1,	B
B	10-17Y	1	N-16, 0-1,	B
B	5-18Y	1	N-10, 0-1,	B
B	20-6Y	1	N-5, 0-5,	B
B	20-31Y	2		B
B	20-39Y	3		B
B	13-25Y	4		B
B	20-1Y	5		B
B	10-20Y	1	N-11, 0-2,	B
B	15-8Y	2		B

80-OTUs: 27·B + 33·HA + 20·M,
Vertical line showed a subject.
Each group showed the subject(s) of a node.
*: "10-11Y"→10-cigarettes/day for 11years
-ested.→ estimated or predicted by DM.
There were also the nodes of non-smokers(A), but lengthy made us to skip them.

of them were quantitatively comprehensible. Moreover, once the structure of the Dt was constructed, as long as the basic concepts of the data were active, all of the following new records could be run on the same Dt. Only with the OTUs data and without the smoking status, the similar identifications are able to build the next feature or attribution of records, which means prediction. Namely, the Dt shown in Fig. 2, is able to classify new data for men and predict who smokes or not.

The main difference between DM and cluster analyses is in how data noise is handled. DM skips noise for a characteristic, e.g., smoking habit, and selects a series of related fields (OTUs); on the other hand, cluster processing respects all data without consideration of any numerical noise.

So, the HIM is known to be individually very different, sensitive and sustainable for long period of time, and will be a source or reservoir of health information that can be evaluated by the application of DM processing.

REFERENCES

1. Jin J, Toyama M, Kibe R, Tanaka Y, Benno Y, Kobayashi T, Shimakawa M, Maruo T, Toda T, Matsuda I, Tagami H, Matsumoto M, Seo G, Sato N, Chounan O, Benno Y. Analysis of the human intestinal microbiota from 92 volunteers after ingestion of identical meals. Benef Microbes 4: 187–193.
2. Sato T, Sato M, Matsuyama J, Kalfas S, Sundqvist G, Hoshino E. 1998. Restriction fragment length polymorphism analysis of 16S rDNA from oral asaccharolytic *Eubacterium* species amplified by polymerase chain reaction. Oral Microbiol Immunol 13: 23–29. [Medline] [CrossRef]
3. Matsuki T, Watanabe K, Fujimoto J, Kado Y, Takada T, Matsumoto K, Tanaka R. 2004. Quantitative PCR with 16S rRNA-gene-targeted species-specific primers for analysis of human intestinal bifidobacteria. Appl Environ Microbiol 70: 167–173. [Medline]
4. Nagashima K, Hisada T, Sato M, Mochizuki J. 2003. Application of new primer-enzyme combinations to terminal restriction fragment length polymorphism profiling of bacterial populations in human feces. Appl Environ Microbiol 69: 1251–1262. [Medline] [CrossRef]
5. Nagashima K, Mochizuki J, Hisada T, Suzuki S, Shimomura K. 2006. Phylogenetic analysis of 16S ribosomal RNA gene sequences from human fecal microbiota and improved utility of terminal restriction fragment length polymorphism profiling. Biosci Microflora 25: 99–107.
6. Matsumoto M, Sakamoto M, Benno Y. 2009. Dynamics of fecal microbiota in hospitalized elderly fed probiotic LKM512 yogurt. Microbiol Immunol 53: 421–432. [Medline] [CrossRef]
7. http://mica.ibest.uidaho.edu/pat.php.
8. Sekirov I, Russell SL, Antunes LC, Finlay BB. 2010. Gut microbiota in health and disease. Physiol Rev 90: 859–904. [Medline] [CrossRef]
9. Santacruz A, Collado MC, GarciaValdes L, Segura MT, MartinLagos JA, Anjos T, MartiRomero M, Lopez RM, Florido J, Campoy C, Sanz Y. 2010. Gut microbiota composition is associated with body weight, weight gain and biochemical parameters in pregnant women. Br J Nutr 104: 83–92. [Medline] [CrossRef]