Comparison of the Accuracy and Mechanism of Data Mining Identification of the Intestinal Microbiota with 7 Restriction Enzymes

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, JLHWO, Kozukue-cho, Kohoku-ku, Yokohama 222-0036, Japan

Received January 15, 2013; Accepted June 6, 2013

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 (OTUs) and subjects’ smoking and drinking habits, which had 2 nominal partitions, yes or no, were analyzed by Data mining software. Identification of subjects for each habit was successfully performed and reported previously, but the identification accuracy was closely dependent on the species of the applied restriction enzymes for PCR. For the sake of better selection of enzymes and understanding the mechanisms of Data mining analysis, 516f-BslI and 516f-HaeIII, 27f-MspI and 27f-AluI and 35f-HhaI, 35f-MspI and 35f-AluI, altogether 7 enzymes, were examined comparatively. Data mining analysis provides a Decision tree for identification of subjects and their dividing pathways that is produced using a limited number of OTUs, which affects the accuracy of the results. The present report discusses not only a global comparison of accuracies for characteristics, but also the detailed mechanisms that result in better or worse results and the practical roles and functions of OTUs. The OTU at the 1st step of the constructed Decision tree was the most important for any identification, and for all cases, the combination of subsequent OTUs, which formed later in the Decision tree, was also unignorable. Detailed dividing pathways were traced and compared for the 7 enzymes and the future supporting ideas were provided for better Data mining analysis of the human intestinal microbiota.

Key words: human intestinal microbiota, restriction enzyme, operational taxonomic unit, smoking and drinking habit, data mining analysis, decision tree, accuracy and mechanism of identification

INTRODUCTION

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 widely performed. In order to analyze and compare the HIM of each subject, various factors, e.g., diets and drugs, that can directly affect the composition of the HIM, must be controlled. In particular, it is essential to unify the dietary factors because daily feeding habits vary among individuals. However, a fixed diet itself may also affect the composition of the HIM, so we cannot apply a fixed diet to long-term feeding experiments. Here we have also tried to apply Data mining analysis (DM) to identify or discriminate the relation between the characteristics of subjects and obtained HIM data from feces. Our previous papers [1, 2] examined how this worked and how we were able to trace the related bacteria with smokers. The obtained results were fruitful, but due them being the first applications of DM analysis to the HIM, deliberate accumulation of a variety of case studies was required to further expand and deepen our ability to stably apply DM. In particular, the selection of primer-restriction enzyme systems (R.Enz.) is the most notable point for practical utilization of DM, because the operational taxonomic unit (OTU) obtained by terminal restriction fragment length polymorphism (T-RFLP) usually contains many different bacteria, and its composition directly affects the results and dictates the accuracy of DM identifications. Regarding the comparison of DM processing with other existing analysis methods, the major difference is that DM utilizes mostly a limited range of OTU data to avoid noises, which are not related to an assigned characteristic.

This paper focuses not only on examination and comparison of practical DM analysis with regard to OTUs but also focuses on increasing the accuracy of
DM identifications. Because DM identifications did not always derive the best or better results, as shown in later tables, much operational experience and many trials were necessary. Here smoking and drinking habits were recognized as examples of a subject’s characteristics. DM analysis is also able to discriminate any OTU data, the characteristic of which have many nominal partitions, but here we applied simple 2 nominal partitions only, because deliberate progress of DM analysis is required for better and stable understanding and utilization.

MATERIALS AND METHODS

To avoid the influences of dietary factors, as already reported [2, 3], we designed identical meals (1,879 kcal/day), which were fed to 92 healthy male volunteers living in Japan for 3 days. The ages and body mass indexes (BMI) of the subjects were 21–59 years (average: 36.8) and 17.3–30.2 kg/m² (average: 22.6), respectively. Fecal samples were analyzed by T-RFLP using 7 R.Enz. [2–4]. 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 subsequent DM analysis, which requires a balance between number of subjects (vertical axis) and the number of OTUs species fields (horizontal axis). In other words, DM analysis requires square or vertically long data, not horizontally long data, so T-RFLP provides an acceptable balance for subjects of 100 or so compared with the other molecular techniques, e.g., metagenomics, which have 1,000 or more species of vertical field. The studies were performed in accordance with the protocol approved by the RIKEN Research Ethics Committee, and the OTUs data were accumulated by the Benno Laboratory, RIKEN, Japan.

Bacterial DNA was isolated from 40–100 mg of feces using the modified method described by Matsuki et al. [5]. Amplification of the fecal 16S rRNA, restriction enzyme digestion, size fractionation of the T-RFs and T-RFLP analysis were carried out as previously described [6–8]. The details of amplification and T-RFLP analysis with the 7 R.Enz., i.e., 516f-BsiI, 516f-HaeIII, 27f-MspI, 27f-AluI, 35f-HhaI, 35f-MspI and 35f-AluI, were described in our previous papers [2, 3].

The amounts for each OTU represent the fluorescence intensity and then concentrations of bacteria. The obtained OTUs data were abbreviated as B— for 516f-BsiI and QA— for 35f-MspI and QA— for 35f-AluI, respectively. We had 2 groups of OTUs: one was 516f- + 27f-, altogether 4 R.Enz., and the other was 35f-, altogether 3 R.Enz. The component numbers of these 7 OTU groups were 27·B, 33·HA, 20·M, 40·A, 31·QA, 34·QM and 48·QA, so if we combined all the enzyme components of each group, the former had a maximum of 120 OTUs, and the latter had a maximum of 113 OTUs. On account of the balance of the number of subjects, i.e., 92, with the number of OTU species field and the problem of field alignment sequence described later, we did not mix the data of the 2 OUT groups. Various sets of R.Enz. combinations were applied, and the data were arranged with the answers of the 92 subjects. The resulting 2-dimensional Excel data were analyzed with DM software (IBM-SPSS, Clementine14).

DM analysis, especially the Classification and Regression Tree (C&RT) DM modeling system, which is the most typical method of subject identification, provides a Decision tree1 (Dt). The Dt identifies explicitly the various groups of subjects according to the assigned characteristic, i.e., smoking [A (No, 76 subjects), B (Yes, 16)] or drinking2 [A (No, 47), B (Yes, 45)] habit in this paper. The C&RT divides subjects into two subsets by comparing the Gini coefficient3 according to the OTUs data, such that the subjects within each subset are more homogeneous than in the previous subset. The C&RT system is quite flexible, and allows unequal misclassification costs to be considered in the other modeling systems of DM. A major speciality of DM and the constructed Dt is that a single selected OUT is used for each step of Dt construction. The default setting of the C&RT system grows a Dt until 5 steps, which is modifiable, e.g., 7 steps, but thinking about the capability of OTUs to discriminate an assigned characteristic, we used 5 steps for the Dt in the present DM identification.

---

1 Decision tree: a decision-supporting pathway that makes use of a tree-like graph.
2 drinking: evening drinking habit, according to the subjects’ answer, refers to normally drinking 1–7 days per week.
3 Gini coefficient: g(t) is well known for quantitative evaluation of the impurity of a group, and is defined at a node t in a C&RT as

\[ g(t) = \sum_{j \neq i} p(j|t)p(i|t) \]

, where i and j are categories of the target field.
RESULTS

Detailed example of a Dt, its construction and accuracy

DM provided a Dt, as shown in Fig. 1, for smoking habit, that discriminates explicitly the various subject groups (i.e., nodes) as boxes. The node at the left end is called the root node in reference to a growing tree, which is the starting point of tree construction. Toward the right side, the Dt grows to divide the subjects appropriately according to the assigned characteristic, i.e., smoking (A: No; B: Yes), with OTUs of 3 R.Enz., i.e., 27·Bsl+33·HaeIII+20·MspI, a total of 80 OTUs in this case.

In Fig. 1, five dotted vertical lines show the growing steps from left to right as Dt 1st step to Dt 5th step, which illustrates the progress of Dt construction. The details of the Dt and the pathway to reach the terminal node⁴ indicated clearly the species and quantities of the related OTUs, which played a role in dividing the various groups of subjects. The Dt also provided practical values of dividing points, that is, the 92 men were divided at Dt 1st step by HA291 into 2 subsets at the left end of Fig. 1 and so on. The critical value for division was 3.13, and at the lower Dt 2nd step, HA291 was also utilized. A major specialty of the C&RT system is that it uses a single selected OTU for each step of the Dt. Seven large arrows indicate terminal nodes containing all 16 smokers (B), and a large dotted arrow indicates a terminal node that contained 56 nonsmokers, i.e., 74% of the ‘A’ group, designated Node-19.

Comparisons between 7 R.Enz.

Table 1 shows a comparison of the results for identification of smoking habit using a single R.Enz. versus combinations of the 7 R.Enz. The upper half of Table 1 shows the 516f- + 27f- group of 4 R.Enz., and the lower half shows the 35f- group of 3 R.Enz. The second row of each group shows the OTU of the Dt 1st step, because this OUT was recognized as having a main role in dividing subjects. The third row for the group indicates the accuracy of DM, i.e., the number of falsely identified subjects in the 92 men until the Dt 5th step, which was the main evaluation term for Dt identification, and the number of falsely identified subjects ranged from 0 to 7.

Fig. 1. Decision tree (Dt) obtained by DM, smoking habit with 80-OTUs: 27·Bsl+33·HaeIII+20·MspI, case 3-A in Table 3.

The large 7 arrows indicate all 16 smokers, and the large dotted arrow indicates gathered 56 nonsmokers. Each box is called a node. The left end, the root node, is the starting point of tree growth toward the right. The name of the OUT, e.g., HA291, that played a role in division is indicated: the numerical dividing point is shown only at Dt 1st step with thin dotted vertical arrow.

---

⁴ Terminal node: tree nodes that do not split further.
Table 1. Smoking habit, 2 Nominal Partitions, accuracy of DM identifications with various restriction enzymes for T-RFLP and their combinations

| species of R.Enz | B | HA | M | A | B + HA | HA + M | B + M | A + B | A + HA | A + M | B + HA + M | HA + M + B | M + B + HA | M + A + HA | A + M + HA | B + HA + M + A |
|------------------|---|----|---|---|--------|--------|--------|-------|--------|-------|-----------|------------|-------------|-------------|-------------|-------------|-------------|
| N. of false identifications in 92 records | 1¢ | 3 | 7 | 4 | 0 | 0 | 1¢ | 1¢ | 1¢ | 4 | 0 | 0 | 0 | 1¢ | 1¢ | 1¢ | 1¢ |

Species of R.Enz QHh QM QA
<table>
<thead>
<tr>
<th>OTU of Dt-1st step</th>
<th>QHh601</th>
<th>QM124</th>
<th>QA829</th>
<th>QHh + QM</th>
<th>QA + QA + QHh</th>
<th>QM+Qh M+QA</th>
<th>QA+QHh</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. of false identifications in 92 records</td>
<td>7</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Abbreviations: Bsl→B (27), HaeIII→HA (33), 27F-Msp1→M (20), 27F-Alu→A (40), 35F-HhaI→QHh (31), 35F-Msp1→QM (34) and 35F-Alu→QA (40); The number in parentheses (n) shows the number of OTUs of each restriction enzyme determined by T-RFLP; Primer restriction-enzymes: R.Enz. The combinations of R.Enz. refer to their sequences during DM processing; N.: number; 'N. of false identifications' means the number of misidentified subjects in the 5th steps of Dt identification by DM: accuracy. *: Figure 1 shows an example of a Dt structure that contains no false identifications; 1¢, 1§, 1※ indicate that the attributed subject is the same; 1£ indicates a subject misidentified only once.

Table 2. Drinking habit, 2 Nominal Partitions, accuracy of DM identifications with various restriction enzymes for T-RFLP and their combinations

| species of R.Enz | B | HA | M | A | B + HA | HA + M | B + M | A + B | A + HA | A + M | B + HA + M | HA + M + B | M + B + HA | M + A + HA | A + M + HA | B + HA + M + A |
|------------------|---|----|---|---|--------|--------|--------|-------|--------|-------|-----------|------------|-------------|-------------|-------------|-------------|-------------|
| N. of false identifications in 92 records | 5 | 2 | 5 | 3 | 5 | 3 | 0 | 2 | 3 | 0 | 1 | 0 | 3 | 0 | 0 | 3 | 0 |

Species of R.Enz QHh QM QA
<table>
<thead>
<tr>
<th>OTU of Dt-1st step</th>
<th>QHh601</th>
<th>QM191Q</th>
<th>QA422</th>
<th>QHh+ QM</th>
<th>QA+QA + QHh</th>
<th>QA+QHh</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. of false identifications in 92 records</td>
<td>12</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Notations are the same with Table 1.

Table 3. Smoking habit, Identifying detailed OTUs for reaching Dt-5th step, 2 Nominal Partitions, representing restriction enzymes and their combinations

<table>
<thead>
<tr>
<th>Case No.</th>
<th>1-A</th>
<th>1-B</th>
<th>2-A</th>
<th>2-B</th>
<th>3-A</th>
<th>3-B</th>
<th>3-C</th>
<th>3-D</th>
<th>3-E</th>
<th>4-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>species of R.Enz</td>
<td>B</td>
<td>HA</td>
<td>B + HA</td>
<td>B + M</td>
<td>B + HA + M</td>
<td>B + HA + M + A</td>
<td>B + HA + M + A + M</td>
<td>B + HA + M + A + M + HA</td>
<td>B + HA + M + A + M + HA + M + A</td>
<td></td>
</tr>
<tr>
<td>OTUs of Dt-4th step</td>
<td>B749, B332, B968, –</td>
<td>HA83, –, HA83</td>
<td>HA83, –, HA83</td>
<td>HA83, –, HA83</td>
<td>HA83, –, HA83</td>
<td>HA83, –, HA83</td>
<td>HA83, –, HA83</td>
<td>HA83, –, HA83</td>
<td>HA83, –, HA83</td>
<td></td>
</tr>
<tr>
<td>N. of false identifications in 92 records</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

*-* means that there was no OUT: in other words, the OTU of one step before is the terminal node; '–(n)' means that the given number of OTUs (n) were continuously absent; Comparison of the underlined OTUs at the Dt 3rd step, comparisons are described in the text; * (3-A): Figure 1 shows an example of a Dt structure; Other notations are the same with Table 1.
A value of 0 represented the best accuracy, and values higher than 0 expressed less accuracy.

In Table 2, the drinking habit of the 92 subjects were identified with their OTUs as shown for smoking in Table 1. Comparing Table 1 and Table 2, the latter reveals a little less accuracy than the former, especially with the three 35f- R.Enz. cases.

**Detailed aspects of Dt construction**

Comparing the various aspects of the obtained Dts, we examined the detailed components of each Dt. Table 3 shows a comparison for smoking habit for cases that were selected from Table 1 based on better accuracy. To assist in understanding the table notation and the Dt pathway, a sample case (case 3-A), which can be traced through Fig. 1, is marked with an asterisk (*) in Table 3. Table 4 shows a similar comparison for drinking habit. As in the case of Fig. 1, to assist in understanding the Dt structure together with the data in the cited table, a sample case (case 2-C’), which is marked (#) in Table 4, is shown in Fig. 2.

**Location of false terminal nodes**

To understand the mechanisms of OTU identification, it was beneficial to not only trace the OTUs that resulted in better accuracy but also those that resulted in worse accuracy. So the location of falsely identified nodes within the Dt were traced, and an example is shown in Fig. 3, which is for drinking and includes 3 misidentified subjects. As easily understood in Table 2 and Table 4, identification was better when the OUT A47 was located at the Dt 1st step. To classify the location of falsely identified nodes, the Dt 1st step was nominated as the main dividing position. The upper half of the Dt was the diluted region (D, Dilute) of A47, the boundary concentration of which was ≤6.65, and the lower half was the concentrated region (C, Conc.) of A47, the boundary concentration of which was 6.65<. The border line is shown as a dotted horizontal line in Fig. 3. The terminal nodes for the false identifications are shown by 3 large dotted arrows. In this case, 2 misleads are situated in Dilute (D2), and 1 is situated in Conc. (C1). In Table 5, which is for both for smoking and drinking habit, typical cases of false locations are shown with emphasis on the misled locations marked with ‘Dilute or Conc.’. A sample case, i.e., Fig. 3, is indicated with a dollar sigh ($) and D2/C1 in the lower part of Table 5.

**DISCUSSION**

**Construction of various Dts and their accuracy**

In Fig. 1, only 8 OTUs out of 80 were active, with 2 OTUs, i.e., HA291 and B749, being applied twice, which indicates that the remaining 72 OTUs were neglected in construction of the Dt. In other words, 8 OTUs were closely related to the subjects’ smoking characteristics, and the other 72 were recognized as OTUs unrelated to smoking like a kind of noise. These facts were the main differences compared with the former classification methods for the HIM, such as clustering, correlation coefficient and principal component analysis (PCA), which consider all OTU data without any selection, and their results are inevitably obscure.

Looking at the lower right side of Fig. 1, terminal nodes were assembled until the Dt 3rd step, and were
not present at the 4th and 5th steps, which meant that
simpler discrimination of OTUs was carried out in this
case. Furthermore, all the terminal nodes in Fig. 1 show
the details of identification of the subjects, and there were
no false identification; that is, the content of all terminal
nodes had 0 at either A or B site. While in some other
cases of Dts shown in Table 1 and Table 2, there were
some false identifications in terminal nodes up to the
Dt 5th step that contained a mixture of ‘A’ and ‘B’ (not
0). These misidentifications were used to evaluate the
accuracy of Dt, which was tightly related to the applied
OTU groups, i.e., species and combinations of R.Enz. So,
this study focused on understanding the mechanisms and
roles of OTUs in order to construct a better Dt structure
with R.Enz.

Accuracies of the 7 R.Enz.

Looking at the cases with better accuracy with regard
to smoking habit in Table 1, the Dt 1st step always had
the same OTU, i.e., HA291 for the 516f- + 27f- group,
and QM124 for the 35f- group, though the latter had less
accuracy than the former. This fact meant that the most
important role, i.e., Dt 1st step, was confirmed for better
identification. Furthermore, applying a larger number
of OTUs, e.g., 4 R.Enz. for the 516f- + 27f- group, did
not always result in better accuracy. On the other hand,
application of a single R.Enz. also showed less accuracy
than a combination of 2 or 3 enzymes. There were 7 cases
of better accuracy, as accuracy=1, these were shown in the
516f- + 27f- group in Table 1. Their constituent subjects
were traced and are indicated with 1$, 1%, 1*, and 1£ in the

Fig. 2. Decision tree (Dt) obtained by DM, drinking habit with 60-OTUs: 20f•27f-MspI+40f•27f-AluI, case 2-C' in Table 4.
The solid arrow indicates the major drinkers group (node), and the 2 dotted arrows indicate the major nondrinkers groups. Other notations
are the same with Fig. 1.
Comparing Table 1 and Table 2, similar features with regard to accuracy were observed. Also, contrasting the 516f- + 27f- group with the 35f- group, the latter had less accuracy than the former. But this was examined only with the two cases, i.e., smoking and drinking habit, so the trend should be confirmed with other cases. Simple imaginations were thought that each OTU element would be dispersed or more uniform with the 35f- group than the 516f- + 27f- group, so clear Dt construction at DM analysis would become rather difficult by the 35f- group.

**Details of better accuracy**

As already described, the CR&T system of DM always divides subjects into two subsets, and principally each Dt step has \(2^{n-1}\) OTUs, where ‘\(n\)’ is the step number, i.e., 1 to 5. So the Dt 5th step fundamentally has 16 OTUs.

---

**Fig. 3.** Decision tree (Dt) obtained by DM, drinking habit with 73-OTUs: 33\( \cdot \)HaeIII+40\( \cdot \)27f\( \cdot \)Alul, marked as $ in Table 5. The 3 dotted arrows indicate the location of falsely identified nodes, ‘D2/C1’. Other notations are the same with Fig. 1.
The accuracies obtained with combinations of OTUs are shown in Table 1 to Table 5.

In almost all of the better cases, except case 1-A in Table 3, the OTU in the Dt 1st step was HA291, and this meant that, as already reported in our previous paper [2], HA291 had a very close relation with smoking in all 4 of these R.Enz., i.e., 120 OTUs. The 4 best cases in Table 3 had no false identifications, i.e., cases 2-A, 3-A, 3-C and 3-D, but the components of their Dt pathways differed slightly between them. Focusing on the Dt 3rd to 5th steps, the components of 2-A and 3-A were completely the same. In other words, although ‘M’ had been applied, the ‘M: 27f-MspI’ of R.Enz. had no effects on 3-A. For 3 cases, 3-A, 3-C and 3-D, we realized that the alignment sequences of the 3 R.Enz. were different and that this had affected greatly the pathway construction. The front alignments in the data sequence were more effective than the back alignments, which is understandable given the fundamental algorithms of the CR&T. At the Dt 3rd step, B749 to M558, B124 to HA175, M133 and B919 to HA83 and M224, (underlined in Table 3) played the same roles in their Dts. A similar trend was also observed with other cases in Table 3. Looking at case 1-B, although we already knew that HA291 was very notable for smoking, a single R.Enz. even ‘HA’ itself, provided less accuracy than a combination of R.Enz. The reason for this was
thought to be follows: even though HA291 was very effective, HA868, in case 1-B at the Dt 2nd step was comparatively less capable of division than B469 and A87 in the same row. Also case 1-B had no Dt 5th step. This indicated interesting mechanisms, that revealed that a single R.Enz. had insufficient OTU data compared with 2 or 3 combined enzymes or that limited species of bacteria had gathered only in HA291 and that other OTUs of ‘HA’ had fewer relations and were insufficient for identification of smoking.

In Table 4 for drinking habit, all the best cases, which had no false identifications, i.e., case 2-A’, 2-C’, 3-B’, 3-C’ and 4-B’, had A47 as their OTU at the Dt 1st step. Concerning their Dt 2nd step, M216 was preferable, but B332 in the case of 2-A’ was replaceable with a combination of OTUs at subsequent steps. In these 5 best cases, none of the cases followed the same pathway before the Dt 3rd step, which seemed to suggest based on comparison that this characteristic, drinking habit, had milder and wider effects on OTUs and more species related to OTUs than smoking. A vigorous OTU like HA291 in Table 3 was not observed in Table 4. In addition, comparing general features between Table 3 and Table 4, the pathways for drinking contained many OTUs at the Dt 5th step and seemed rather packed compared with the pathways for smoking. Generally, if the number of OTUs utilized in Table 3 and Table 4 is considered, the pathways for drinking contained more OTUs, i.e., more related OTUs species, than the pathways for smoking.

Detailed aspects of worse accuracy

In Table 5, the location of falsely identified nodes in the Dt, whether in the ‘Dilute or Concentrated region’ shown in Fig. 3, revealed the features of misleading mechanisms. Most of false location (around 3/4) were situated in Dilute and separated by the Dt 1st step, and this indicated that identification was difficult for them on account of the signatures of the OTUs, for smoking or drinking, being too ambiguous or obscured to distinguish. On the other hand, around 1/4 of the false locations were found in the Concentrated region (Conc.) in Table 5, and this meant that other mechanisms misled pathways compared with those in Dilute. This is, the subsequent OTU species after the Dt 2nd step were not able to distinguish properly due to slight differences in the fitted OTUs in Conc. While, the 2 rare cases of ‘D2/C1’ in Table 5, observed only at drinking table, indicated as the mixed mechanisms described above.

Furthermore, Table 5 gave us some additional estimations, that the related OTUs, i.e., group of bacteria, to an assigned characteristic, did not originally exist so much, therefore how the limited species of bacteria were belonged to a certain OTUs were the key and mechanism of these DM identifications, which were also recognized in Table 3 and Table 4. On the other hand, if a plentiful number of related bacteria species did exist, not limited to a small amount, then the locations of false terminal nodes like in Table 5 would be more spread out and blended.

Finally, the major results of applying DM analysis to the HIM were identification of subjects with only several closely related OTUs, i.e., groups of bacteria, and the finding of explicit and numerically clear interdependences between the assigned characteristic and OTUs. These were the most remarkable differences compared with the former classification methods, like clustering, correlation coefficients and PCA. The constructed Dt ignored most of the less related OTUs, treating them like noise, which was not possible with any of the former methods. However as already described, identification of OTUs for assigned characteristics was not easily performed with a Dt. Precise and thoughtful preparations are necessary, and in particular, the most important step in preparation is the selection of R.Enz. species. Here we applied only 2 characteristics, smoking and drinking, so recommendations for some sets of R.Enz. species could not be made at this time, but based on Table 1 to Table 5, some advice can be proffered. That is, 2 to 3 sets of R.Enz. species were eligible than single or 4 R.Enz. for 2 nominal partitions of characteristic and for subjects of 100 or so.

To obtain clear identifications or the best accuracy, the OTU at the Dt 1st step played the most important role, and the reason why an OTU, e.g., HA291 for smoking, revealed such an activity depended on each characteristic and specialty of the R.Enz., which should be clarified by accumulation of similar analyses and experience in the future. However, it was certain that a combination of several OTUs at subsequent steps of the Dt was required to obtain the best accuracy, actual examples of which are shown in Table 3 and Table 4.

As for the OTU species that were situated at the Dt 4th to 5th steps, which were surely less related to the assigned characteristics than those situated at the Dt 1st to 3rd steps, there was some slight doubt concerning the need for further detailed tracing. Thinking about the accuracy of identification with various OTU data, the fundamental algorithms of DM software and unknown symbiotic characters of various uncultured bacteria, which have already been reported [2], this was not something easy to resolve. Some intuitive and practical streamlining focused on clinical applications would be
preferable to strict adherence to accuracy with regard to subjects who are borderline. The main reasons were as follows. First, the OTUs themselves were not absolutely firm and stable, and would be affected by peripheral circumstances, e.g., latest meals and drugs, personal factors and living localities of the subjects. Second, the assigned characteristics, smoking or drinking, were not always strictly defined, and so it was possible that they were defined based on personal concepts and that they contained wide intermediate stages. Third, thinking about the future application of Dts for predictive analysis of diseases, such as alimentary disorders, it is expected that identifications and predictions derived from a Dt would support diagnoses and be sufficient to suggest other possibilities.

Once a certain Dt is constructed with a well-considered group of subjects, a subsequent new group of similar subjects for which there is less clinical information can be run on the same Dt, and their possibilities of suffering similar disorders can be identified, which will be very effective for new forms of preventive medicine. The HIM is known to be different between individuals, very sensitive and closely related to various physiological characteristics, suggesting that the HIM and OTUs can become a new source or reservoir of health information that can be used to evaluate patients and make predictions using the Dt structure constructed with DM analysis.

ACKNOWLEDGEMENTS

All data processed here were accumulated cooperatively with Jong-Sik Jin, Mutsumi Touyama, Ryoko Kibe, Yoshiki Tanaka, Yoshiko Benno and Yoshi Benno, RIKEN, and we appreciate their collaborative efforts. In addition, we are also grateful to Saburou Nakazawa, MD, PhD, the former General Director of the Japanese Society of Gastroenterology, for his excellent and insightful encouragement of these studies, and appreciate the steady and long-term financial support provided by Tatsuo Shimamoto, MD, PhD, and Mr. Toshiaki Teramura, Arteriosclerosis Research Foundation, Japan, for DM facilities and operations.

REFERENCES