Visualizing Individual and Region-specific Microbial–metabolite Relations by Important Variable Selection Using Machine Learning Approaches

Satoshi Tsutsui¹, Yasuhiro Date¹,², and Jun Kikuchi¹,²,³*

¹ Graduate School of Medical Life Science, Yokohama City University, 1-7-29 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan.

² RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan.

³ Graduate School of Bioagricultural Sciences, Nagoya University, 1 Furo-cho, Chikusa-ku, Nagoya, Aichi 464-0810, Japan.

(Received June 1, 2017; Accepted June 20, 2017)

Data mining techniques such as machine learning have greatly advanced the chemical and biological sciences. Especially, technological advances in data mining are anticipated for analyzing big data derived from biological and environmental systems. From this perspective, we analyze the complex metabolic and microbial responses of human skin and the relations among these responses using advanced data mining techniques. To this end, metabolic profiles of human sweats were characterized via multiple NMR spectra, followed by an advanced analytical strategy based on data-driven and machine learning approaches. These methods extracted the important variables of the metabolites associated with microbial community variations. Moreover, the relation between the sweat metabolites and the skin microbes was successfully visualized by correlation-based networks. This analytical strategy promises a versatile and useful approach for big data analyses in various fields of science.

Key Words: Human sweat, Metabolic profiling, NMR, Skin microbiota, Machine learning, Important variables, Data science

Introduction

The remarkable development of computer-based technologies has enabled technological innovations in multiple scientific disciplines. Particularly, computing
power has enormously increased and current processing speeds can handle large amounts of information, enabling high-speed processing for elaborate calculations such as deep learning\(^1\). Machine learning (which is typified by deep learning) has also tremendously diversified in recent years. Machine learning approaches are applied in big data analyses in various scientific fields, including the biological sciences.

Machine learning approaches are also utilized in comprehensive analyses of chemical compounds and metabolites in complex mixtures derived from biological systems. In comprehensive metabolic profiling, the data are acquired by measurement instruments such as nuclear magnetic resonance (NMR) spectroscopy\(^2\) and analyzed by machine learning and multivariate analyses. The NMR-based approach is applicable to various types of samples, including intact cells and tissues\(^3\)-\(^6\) and mixtures in structural studies\(^7\)-\(^9\). NMR also provides accurate quantification\(^10\) with inter-institution convertibility\(^11\). In previous studies, NMR-based metabolic profiling has been applied to plants\(^12\)-\(^15\), animals\(^16\)-\(^20\), and microbial\(^21\)-\(^24\) systems.

Humans are the commonest and most important research subjects in NMR-based metabolic profiling. From an ethical perspective, non-invasive samples such as urine, feces, and saliva are preferable for the analysis in NMR-based metabolic profiling. However, NMR-based metabolic profiling has rarely targeted human sweat despite the relative ease of collecting sweat samples. Thus, very little information on sweat metabolic profiles (NMR spectra) and sampling procedures is available in previous reports\(^25\)-\(^26\). In contrast, quantitative and qualitative information on other human samples (e.g., urine\(^27\), serum\(^28\), and saliva\(^29\)) is readily accessible.

Annotation and identification of metabolites in complex chemical mixtures is an important step in NMR-based metabolic profiling of human samples. This is usually accomplished by multidimensional NMR measurements, database searching, statistical techniques, and by searching other information published in scientific journals and assisted by other instrumental data. Metabolite databases and assignment tools are especially valuable for metabolite characterization in NMR-based metabolic profiling. Such databases and tools are publically available on the web or as software. To complement public databases such as the Human Metabolome Database\(^30\) and the Biological Magnetic Resonance Data Bank\(^31\), we have developed a number of assistance tools (databases) for metabolite annotations, namely SpinAssign\(^32\)-\(^33\), SpinCouple\(^34\), and signal enhancement by spectral integration (SENSI) programs\(^35\)-\(^36\). The SpinAssign and SpinCouple are web-based programs for metabolite annotation accomplished by database searching in \(^1\)H--\(^1\)H heteronuclear single quantum coherence (HSQC) or two-dimensional (2D) \(^1\)H--\(^1\)H \(J\)-resolved (\(J\)res) NMR data depositories, respectively, and SENSI is a statistical program for metabolite annotation assisted by computational improvement of the signal-to-noise ratio. These programs can assist the metabolite annotations and assignments of human sweat samples.

This study characterizes the metabolic mixtures in human sweat samples by NMR-based metabolic profiling combined with database searching methods such as SpinAssign and SpinCouple. The metabolites in human-produced sweat, which are co-modulated by microbial symbions on human skin, are characterized by the NMR-based approach. We also analyze the microbial communities on skin and their relations with sweat metabolites. The relations between microbial community variations (in different localizations) and metabolic variabilities cannot be elucidated by conventional approaches such as principal component analysis (PCA) and correlation-based analysis. For this reason, we propose an analytical strategy that identifies the important variables (metabolites) associated with microbial community variations by data-driven and machine learning approaches.

## Materials and Methods

### Sweat samples

The analysis of sweat samples from human volunteers was approved by the human ethical committees of RIKEN and Yokohama City University. All volunteers (\(n = 6\)) provided informed consent prior to the study. The sweat samples were collected from the face, neck, arm, armpit, breast, back, buttock, and leg using cellulose materials (Kimwipes (Nippon Paper Crecia Co. Ltd., Tokyo, Japan) for metabolic analysis and swabs for microbial analysis). A total of 111 samples from six volunteers and 72 samples from one volunteer were collected for metabolic and microbial characterization, respectively. The sweat samples were stored at \(-20^\circ\text{C}\) until required for further analyses.

### NMR measurements

The collected sweat samples adsorbed to the cellulose materials were added to a 15-ml tube containing 2 ml phosphate buffer (90\% \(\text{D}_2\text{O}\), 10 mM DSS, 0.1 M \(\text{K}_3\text{HPO}_4/\text{KH}_2\text{PO}_4\), pH 7.0), and incubated with shaking (750 rpm) for 15 min at 65°C. The incubated samples were then centrifuged at 6000 rpm for 10 min at 4°C, and 600 \(\mu\)l of the effluent was collected and transferred to an NMR tube. All NMR measurements were performed in an NMR machine (AVANCE II 700, Bruker, Billerica, MA, USA) with a \(1\)H inverse triple-resonance cryoprobe (\(1\)H: 700.153 MHz, \(^1\)C: 176.061 MHz). The NMR measurement procedures are detailed in previous
papers . Briefly, the 1H NMR measurements were performed by the nuclear Overhauser effect correlated spectroscopy (NOESY) pulse program, operated at 298 K with the following parameters: 65536 data points, 14098 Hz spectral width, 3 s relaxation time, 10 ms mixing time, number of dummy scans = 4, and number of scans = 256.

For metabolite annotations, 6 ml of each sweat sample solution was evaporated and condensed in a vacuum freeze dryer. The condensed sample was redissolved in 1.6 ml phosphate buffer, and the sample solution was transferred to an NMR tube. The metabolites in the condensed sweat sample were annotated by 1H–1H total correlation spectroscopy (TOCSY), 13C–1H HSQC, and 2D 1H–1H Jres NMR measurements. The parameter settings were 512 (f1) and 2,048 (f2) data points, spectral widths 7,003 Hz (f1) and 14,098 Hz (f2), relaxation time = 3 s, mixing time = 70 ms, number of dummy scans = 4, number of scans = 256 for the 1H–1H TOCSY NMR measurements, 256 (f1) and 2,048 (f2) data points, 256 (f1) and 2,048 (f2) data points, spectral widths 50 Hz (f1) and 14,098 Hz (f2), number of scans = 8 for the 2D Jres NMR measurements. The data were processed by the TopSpin software package (Bruker Biospin, Rheinstetten, Germany). The metabolite annotations were assisted by SENSI, the SpinAssign program on the PRIME website, the SpinCouple program, the Human Metabolome Database, and by references to previously published data.

**Microbial community analysis**

DNA was extracted from the collected sweat samples by a QIAamp DNA Mini Kit (QIAGEN), following the manufacturer’s instructions. The DNA amplification and sequencing methods are described in a previous study. Briefly, the extracted DNA was amplified by the polymerase chain reaction using TaKaRa Ex Taq (TAKARA Bio Inc., Tokyo, Japan) as follows: an initial denaturing step at 95°C for 4 min, followed by 35 denaturation cycles at 95°C for 30 s, annealing at 55°C for 30 s, extension at 72°C for 1 min, and a final extension step at 72°C for 10 min. The primers used in this study are given in Kozich et al. The sequencing was performed on a MiSeq sequencer (Illumina Inc., San Diego, CA, USA) following the manufacturer’s instructions. The obtained sequencing data were analyzed on QIIME.

---

**Figure 1** 1D and 2D NMR spectra for identifying the metabolites in human sweat samples. The displayed spectra were obtained from SENSI based on 1H NOESY spectra (A), 1H–1H Jres NMR spectrum (B), 1H–1H TOCSY spectrum (C), and 13C–1H HSQC spectrum (D).
Data analysis

The obtained NMR data were normalized by probabilistic quotient normalization\(^{42}\) in the mQTL package of the free software R (https://www.r-project.org/), then aligned by icoshift\(^{43}\) in MATLAB software (MathWorks Japan, Tokyo, Japan). The signal peak detection and noise eliminations in the aligned spectra were subsequently performed on the SENSI 1D web site (http://emar.riken.jp:3838/SENSI_1D/). The processed metabolic and microbial data were analyzed by PCA and correlation analysis in R software and Microsoft Excel as described in a previous study\(^{44}\). The data from the correlation analysis were depicted as a network diagram by gephi (https://gephi.org/) as previously described \(^{45,46}\). Random forest\(^{47}\) conditional variable importance measures (cforest)\(^{48}\) and the support vector machine (SVM) were implemented in the “party”\(^{49}\) and “classyfire”\(^{50}\) packages, respectively, and run in R software. The models constructed by cforest were verified by leave-one-out cross validation.

Results and Discussion

This study characterized the metabolites in human sweat samples by combining multiple NMR measurements with database searching and the SENSI method. Subsequently, an analytical strategy that selects the important variables was proposed, and the metabolic and microbial variations and their relations were evaluated.

Characterization of sweat metabolites

Unlike the metabolites in human urine and serum, the complex metabolite mixtures in human sweat have rarely been characterized by NMR-based approaches. In a preliminary test of the sampling materials, Kimwipes was the best cellulosic material for characterizing the sweat metabolites by NMR. The other tested materials were cotton, Salivette (Sarstedt, Tokyo, Japan), and absorbent polymers (data not shown). The spectral profiles differed between the eccrine gland derived from the arm and the apocrine gland derived from the armpit (Fig. S1). The sweat samples from 6 volunteers were collected by an optimized sampling procedure. The sweat metabolites were annotated and identified in 1D and 2D NMR measurements, namely, by H NOESY, \(^1\)H–\(^1\)H TOCSY, \(^1\)C–\(^1\)H HSQC, and \(^1\)H–\(^1\)H Jres NMR measurements. In addition, the \(^1\)H NOESY spectra were analyzed by SENSI (Fig. 1). The obtained spectra were further analyzed by database searching, assisted by the metabolite annotation tools SpinAssign and SpinCouple. The final annotated and assigned metabolites in human sweat samples are listed in Table 1.

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>(^1)H Chemical Shift (ppm) and multiplicity</th>
<th>NMR spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taurine</td>
<td>3.44 ppm (6H)</td>
<td>1D</td>
</tr>
<tr>
<td>Proline</td>
<td>7.20 ppm (2H, 7.53 ppm (2H, 7.33 ppm (2H, 4.20 ppm (4H, 3.51 ppm (8H, 3.34 ppm (4H, 2.53 ppm (4H, 1.97 ppm (8H, 1.36 ppm (4H, 1.25 ppm (4H, 0.95 ppm (12H, 0.83 ppm (12H)</td>
<td>2D</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>7.40 ppm (2H, 7.34 ppm (2H, 6.87 ppm (4H, 2.05 ppm (4H, 1.34 ppm (4H, 1.24 ppm (4H, 0.85 ppm (12H, 0.78 ppm (12H)</td>
<td>1D</td>
</tr>
<tr>
<td>Valine</td>
<td>7.00 ppm (2H, 6.97 ppm (2H)</td>
<td>1D</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>8.08 ppm (2H, 8.06 ppm (2H)</td>
<td>1D</td>
</tr>
<tr>
<td>Histidine</td>
<td>7.10 ppm (3H, 4.01 ppm (3H, 3.29 ppm (3H, 1.88 ppm (3H, 1.76 ppm (3H, 1.35 ppm (3H, 1.22 ppm (3H, 0.87 ppm (3H, 0.78 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>Proline</td>
<td>8.08 ppm (2H, 8.06 ppm (2H)</td>
<td>1D</td>
</tr>
<tr>
<td>Glu</td>
<td>2.55 ppm (4H)</td>
<td>1D</td>
</tr>
<tr>
<td>Serine</td>
<td>4.05 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>Asp</td>
<td>4.10 ppm (1H)</td>
<td>2D</td>
</tr>
<tr>
<td>Lys</td>
<td>1.80 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>Gly</td>
<td>5.06 ppm (1H)</td>
<td>10</td>
</tr>
<tr>
<td>Cys</td>
<td>3.75 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>SArg</td>
<td>3.58 ppm (2H, 2.87 ppm (2H, 1.47 ppm (4H)</td>
<td>1D</td>
</tr>
<tr>
<td>Arginine</td>
<td>3.77 ppm (2H, 4.01 ppm (4H, 3.14 ppm (4H)</td>
<td>1D</td>
</tr>
<tr>
<td>Ornithine</td>
<td>3.77 ppm (2H, 3.62 ppm (3H, 1.16 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>Citrulline</td>
<td>3.75 ppm (2H, 3.64 ppm (2H, 1.84 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>Arginine</td>
<td>3.75 ppm (2H, 3.64 ppm (2H, 1.84 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Citrulline</td>
<td>3.77 ppm (2H, 3.62 ppm (3H, 1.16 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>Glutamine</td>
<td>3.75 ppm (2H, 3.64 ppm (2H, 1.84 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>Glutamine</td>
<td>2.25 ppm (2H, 2.25 ppm (2H, 1.73 ppm (2H)</td>
<td>2D</td>
</tr>
<tr>
<td>Amines</td>
<td>3.56 ppm (3H, 1.01 ppm (3H, 0.55 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>Ethanol</td>
<td>3.94 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>Valine</td>
<td>3.90 ppm (1H, 2.38 ppm (2H, 1.03 ppm (4H)</td>
<td>1D</td>
</tr>
<tr>
<td>Glutamine</td>
<td>2.55 ppm (1H)</td>
<td>2D</td>
</tr>
<tr>
<td>L(+)-Tartaric acid</td>
<td>3.54 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Leucine</td>
<td>3.58 ppm (2H, 2.05 ppm (2H, 1.09 ppm (2H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Lysine</td>
<td>3.56 ppm (3H, 1.05 ppm (3H, 0.54 ppm (3H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Arginine</td>
<td>2.90 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Tyrosine</td>
<td>2.70 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Isoleucine</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Threonine</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Valine</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Proline</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Histidine</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Proline</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Arginine</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Valine</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Asparagine</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Asparagine</td>
<td>2.76 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>Acetate</td>
<td>1.95 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Malic acid</td>
<td>1.95 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>L-Glutaric acid</td>
<td>1.95 ppm (1H)</td>
<td>1D</td>
</tr>
<tr>
<td>Lipid (free)</td>
<td>0.37 ppm (1H)</td>
<td>1D</td>
</tr>
</tbody>
</table>

Metabolic and microbial analyses

To discern the trends in the datasets derived from human sweat samples, the obtained \(^1\)H NMR spectra were analyzed by an unsupervised multivariate analysis technique (namely, PCA). The metabolic profiles were not clustered based on skin localization, individuals, circadian rhythms, and day-to-day variations in the PCA scores plot (Fig. 2A). In contrast, the microbial community analysis by the Miseq sequencer revealed dominant bacterial phyla such as Proteobacteria and Actinobacteria on each body part (Fig. S2). In the PCA-based microbial profiling, locality on the human skin predominantly determined the profiles on the PCA scores plot (Fig. 2B). Similar results were obtained in the correlation analysis (Fig. S3), and in a previous report\(^{51}\), i.e., the microbial communities varied by habitat (body part) on human skin. The PCA results also clustered the microbial profiles into two groups; the face, neck, and armpit group and the arm, leg, and buttock group (Figs. 2B and S4). However, conventional approaches such as PCA and correlation analysis cannot capture the metabolic features of human sweats. Therefore, they cannot reveal the relations between the sweat metabolites...
and skin microbial communities. For this purpose, an analytical strategy that identifies the important variables (metabolites) associated with the microbial community profiles is required.

![Fig. 2](image-url) Metabolic (A) and microbial community (B) profiles evaluated by PCA. Red circles, triangles, and squares indicate arm, buttock, and leg samples, respectively. Black circles, triangles, and squares indicate face, neck, and armpit samples, respectively.

![Fig. 3](image-url) Strategy for selecting the important metabolite variables associated with the microbial community profiles.

![Fig. 4](image-url) Most important sets of variables screened by the RIIVS criterion. Shown are the screenings from the top 40 to the top 5 (A) and verification of the selected important variables (from 9 to 5) (B).

**Selection of important variables**

The microbial community analysis roughly clustered the PCA profiles into two groups of body parts. This two-group categorization of the microbial profiles was applied to the metabolic data analysis. In other words, we applied the class information of the microbial profiles to the further analysis of sweat metabolites as a data-driven approach. The important variables (metabolites) that significantly contribute to the differences between the two groups were then determined by machine learning approaches. In the proposed analytical strategy, the important variables were selected by two machine learning algorithms, cforest and SVM (Fig. 3). The cforest algorithm selected the important variables, and the SVM algorithm validated the selected variables. The variables of the metabolites ranked by cforest were aligned in order of importance (Fig. S5), and numbered from top down (from the top 5 to 40) for validating the SVM (Fig. 3). In the SVM validations, the improved classification performance was standardized as a relative indicator for important variable selection (RIIVS) as follows:

\[
\text{RIIVS} = \frac{\text{CP}_1 - \text{CP}_0}{100 - \text{CP}_0}, \quad (1)
\]

where CP is the classification performance (%), and CP\(_0\) and CP\(_1\) are the classification performances of the SVM using all variables and the selected variables, respectively. After screening the most important sets of variables by this RIIVS criterion (Fig. 4A), six metabolic variables were identified as the major contributors to the categorization of body parts (Fig. 4B). The selected variables, which annotated as 1,2-propanediol, ethanol, and unknown metabolites (Fig. 5), were subsequently verified by PCA. By this analysis, the metabolic profiles were clustered into two localizations of body parts (Fig. 6A). Thus, the adopted analytical strategy was successfully improved by combining the two machine learning algorithms (cforest and SVM). Applying this analytical strategy to the microbial data, nine microbial species were selected as the important variables. Collectively, the six metabolites and nine metabolites improved the clustering performance of PCA, as shown in Fig. 6B.

**Metabolic and microbial relations**

The relation between the sweat metabolites and skin microbial communities was elucidated by a correlation-based analysis of the six metabolites and nine microbial species selected by the proposed analytical strategy (Fig. 7). As shown in the correlation network depicted in gephi, the important metabolites are correlated with certain microbial species (e.g., members of Propionibacteriaceae and Corynebacteriaceae).
Fig. 5 Comparison of 1D $^1$H NMR spectrum (A) and 2D Jres projection spectrum (B). Arrows indicate the selected important metabolites annotated as 1,2-propanediol (d), ethanol (e), and unknown metabolites (a, b, c, and f).

Fig. 6 Evaluation of clustering performance on PCA scores plots based on the selected metabolites (A) and on both microbes (nine variables) and metabolites (six variables) (B).

Therefore, the proposed analytical strategy can capture the relations between the sweat metabolites and skin microbial communities. In a wider sense, this strategy (which combines cforest screening with SVM validation) will assist researchers in determining the most important sets of variables contributing to a categorization. This analytical strategy is applicable not only to metabolic and microbial data, but also to the other big data obtained from a variety of measurement instruments.

Conclusions

This study characterized the metabolic profiles of human sweat and evaluated the relations between the metabolic variations and microbial community profiles. This was achieved by optimizing the sampling procedure and acquiring multiple NMR spectra for annotating and identifying the sweat metabolites. In addition, an analytical strategy that selects the important variables contributing to a categorization was proposed. The selection is made by combined cforest screening and SVM validation under a RIIVS criterion. Using a data-driven approach, the relation between sweat metabolites and skin microbial species was visualized in correlation-based networks of the important variables selected by the analytical strategy. This analytical strategy should provide a versatile and useful approach for big data analyses in various scientific fields.

Fig. 7 Network diagrams based on correlation analysis of metabolites and microbes. Shown are the relations between the selected metabolites and the selected microbes. Relatively strong, middle, and weak correlations are rendered in red, orange, and green, respectively. a, b, c, and f, unknown metabolites; d, 1,2-propanediol; e, ethanol.

Acknowledgments

The authors wish to thank Takuma Misawa, Taiga Asakura, Kengo Ito, and Yuka Shiokawa (Yokohama City University) for useful advices and discussions on sampling, NMR measurements, Miseq sequencing, and data mining. This research was supported in part by a Grant-in-Aid for Scientific Research (to J.K.). The authors would like to thank Enago (www.enago.jp) for the English language review.

References and Notes


Supplementary Materials

Fig. S1 ¹H NMR spectra of eccrine gland derived from the arm and apocrine gland derived from the armpit.

Fig. S2 Microbial compositions detected on different body parts.
Fig. S3 Correlation heat map based on the microbial community profiles leg, and buttock group.

Fig. S4 Microbial community differences between the face, neck, and armpit group and the arm, leg, and buttock group.
Fig. S5 Important variables characterized by the cforest approach leg, and buttock group.