SCIENTIFIC REPORTS

OPEN

Received: 28 April 2017
Accepted: 8 January 2018
Published online: 21 February 2018

Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet

Yuka Shiokawa[1,2], Yasuhiro Date[1,2] & Jun Kikuchi[1,2,3]

Computer-based technological innovation provides advancements in sophisticated and diverse analytical instruments, enabling massive amounts of data collection with relative ease. This is accompanied by a fast-growing demand for technological progress in data mining methods for analysis of big data derived from chemical and biological systems. From this perspective, use of a general "linear" multivariate analysis alone limits interpretations due to "non-linear" variations in metabolic data from living organisms. Here we describe a kernel principal component analysis (KPCA)-incorporated analytical approach for extracting useful information from metabolic profiling data. To overcome the limitation of important variable (metabolite) determinations, we incorporated a random forest conditional variable importance measure into our KPCA-based analytical approach to demonstrate the relative importance of metabolites. Using a market basket analysis, hippurate, the most important variable detected in the importance measure, was associated with high levels of some vitamins and minerals present in foods eaten the previous day, suggesting a relationship between increased hippurate and intake of a wide variety of vegetables and fruits. Therefore, the KPCA-incorporated analytical approach described herein enabled us to capture input–output responses, and should be useful not only for metabolic profiling but also for profiling in other areas of biological and environmental systems.

Innovation in computer-based technology has caused not just advancements of computer-associated technology but also considerably contributions of their ripple effects to technological progress in research fields of chemistry and biology. This technological innovation facilitates advancements in sophisticated and diverse analytical instruments, enabling massive amounts of data collection with relative ease. The increasing opportunity of handling "big data" has accompanied with a fast-growing demand for technological progress in highly analytical methods for mining "big data." From this viewpoint, machine learning approaches such as deep learning and data mining techniques are currently being developing at a fast clip.

One research field in chemistry and biology that acquires and handles "big data" is metabolomics or metabolic profiling. A massive amount of data in metabolomics studies is typically obtained by nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry. Especially, NMR spectroscopy is a non-destructive method for measurements of complex metabolites derived from biological systems[1–3]. In addition, NMR has advantages for analytical reproducibility and inter-convertibility among different institutions[4,5]. Therefore, NMR-based metabolic profiling has been applied to various biological and environmental samples[6–13]. These types of research benefit from several useful and helpful databases and analytical support tools for preprocessing of spectral data and assignments of metabolites in complex chemical mixtures in NMR-based metabolic profiling. Such databases

[1]RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 235-0045, Japan. [2]Graduate School of Medical Life Science, Yokohama City University, 1-7-29 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan. [3]Graduate School of Bioagricultural Sciences and School of Agricultural Sciences, Nagoya University, 1 Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan. Correspondence and requests for materials should be addressed to J.K. (email: jun.kikuchi@riken.jp)

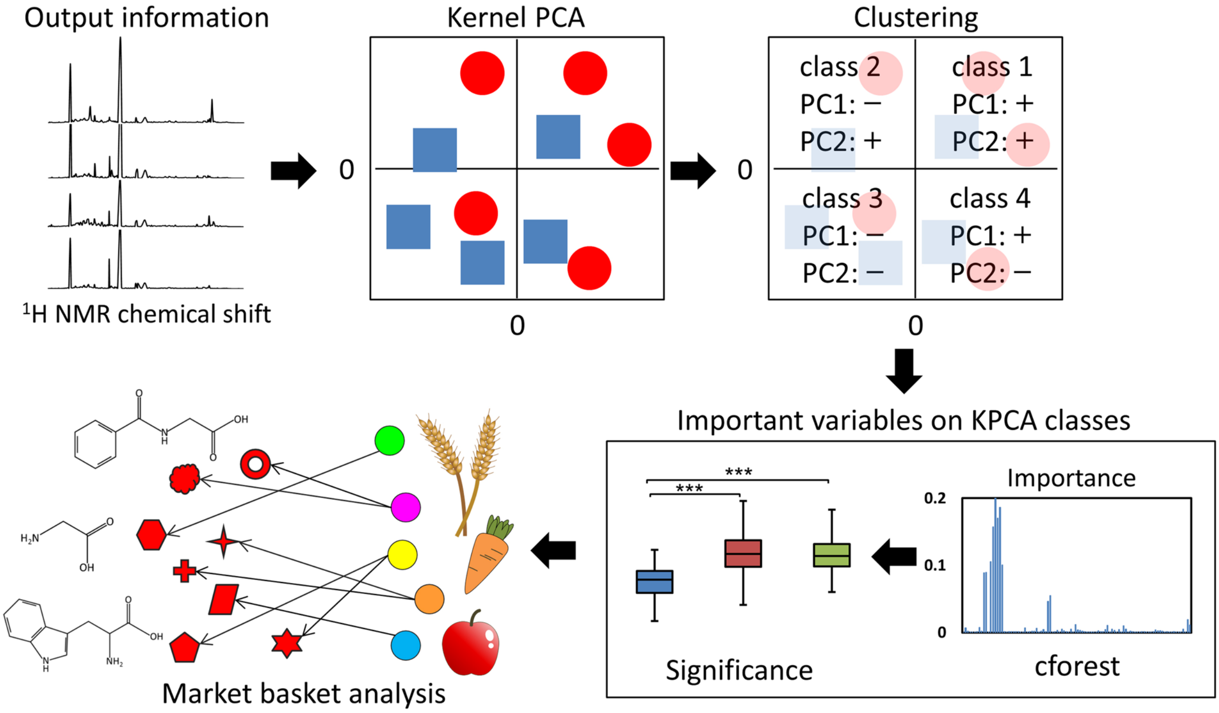

Figure 1. Analytical flow of the study. Kernel principal component analysis (KPCA) was calculated from urinary organic (nuclear magnetic resonance) and inorganic (inductively coupled plasma optical emission spectrometry) data. Subsequently, cforest was used to identify significant metabolites in four groups generated from the KPCA results. Finally, market basket analysis was used to obtain human lifestyle-associated information related to significant metabolites. The colored circles and red symbols in the market basket analysis indicate individual nutrients and metabolites (for the purposes of illustration), respectively. The image was drawn by Yuka Shiokawa.

and tools, for example, include the human metabolome database[14], biological magnetic resonance data bank[15], ^{1}H (^{13}C) TOCCATA[16,17], SpinAssign[18], SpinCouple[19], MetaboAnalyst[20], NMRShiftDB[21], MVAPACK[22], rNMR[23], BATMAN[24], statistical total correlation spectroscopy[25], fragment-assembly approach[26], and signal enhancement by spectral integration (SENSI) method[27].

In the fields of NMR-based metabolomics, one key multivariate analysis is principal component analysis (PCA). PCA is an unsupervised method and a kind of "linear" multivariate analyses. Although PCA is able to capture meaningful tendencies in some datasets, PCA is not a panacea for analyses in all instances. For example, some cases in metabolic variations have typically non-linear relationships. Therefore, general PCA is not able to capture "non-linear" metabolomic relationships from various metabolic reactions in living organisms. Therefore, advances in powerful data mining methods that would allow the discoveries of valuable information from massive datasets have been eagerly anticipated. To circumvent difficulties in obtaining valuable information that cannot be extracted by conventional linear PCA methods, we focused on "non-linear" kernel PCA (KPCA)[28]. KPCA is an enhanced PCA method that incorporates a kernel function, thereby facilitating solution of non-linear problems. KPCA was previously applied to analysis of NMR-based metabolic profiling[29]. However, KPCA is limited by an inability to determine importance of variables in contrast to linear PCA where it is possible to identify key variables that contribute to PCA score profiles. Thus, it was important to overcome this limitation for evaluation of key metabolites and for discovery of useful biomarkers in NMR-based metabolic profiling studies. To identify key variables for kernel-based methods, several variable selection approaches have been previously reported in supervised classifications and regressions[30,31], however, determination of important variables for unsupervised data using kernel-based methods is still challenging.

Here, we describe a KPCA-incorporated analytical approach for the extraction of useful information from NMR-based metabolic profiling datasets. To overcome the limitation concerning important variable identifications in unsupervised KPCA, we incorporated a random forest conditional variable importance measure (cforest)[32], a form of machine learning, into the KPCA-based analytical approach to determine the importance of variables. The obtained importance was validated using statistical tests and further analyzed using a market basket analysis (MBA)[33] to evaluate input–output responses (urinary metabolites and minerals associated with dietary food and nutritional information) in humans (Fig. 1).

Results and Discussion

Non-linear KPCA. In this study, urinary metabolic and elemental data obtained from NMR and inductively coupled plasma optical emission spectrometry (ICP-OES), respectively (Figure S1), were integrated on a data matrix prior to KPCA. KPCA was performed using the analysis of variance (ANOVA) kernel function after

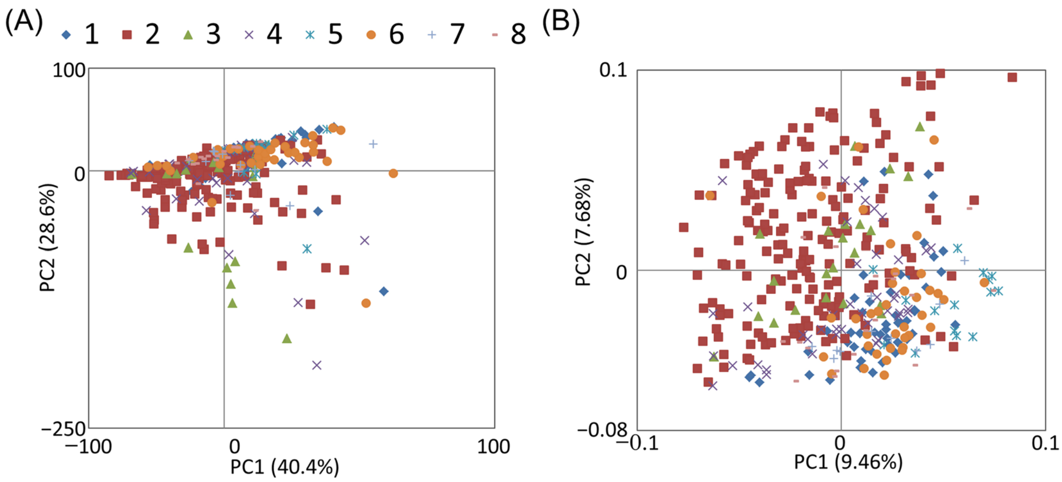

Figure 2. Comparison of profiles from conventional principal component analysis (PCA) (**A**) and kernel PCA (**B**). The symbols and numbers indicate individual subjects.

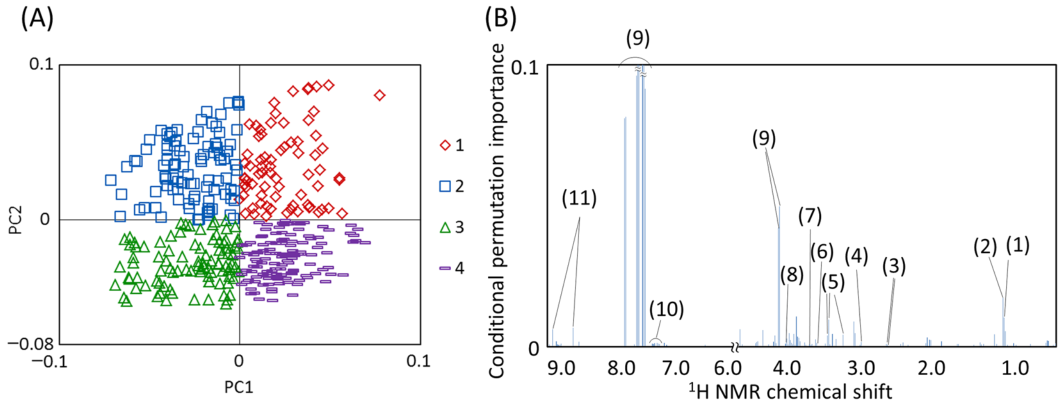

Figure 3. Important variables evaluated in the cforest analysis. Kernel principal component analysis results were used to generate four groups based on PC1 and PC2 plus and minus signs for the cforest analysis (**A**). The top 100 variables were depicted as important variables determined by cforest (**B**). The above-highlighted numbers correspond to metabolites as follows: 1: Methylmalonate, 2: Fucose, 3: Citrate, 4: Creatinine, 5: Taurine, 6: Glycine, 7: Ethanol, 8: Glucose, 9: Hippurate, 10: Phenylalanine, 11: Trigonelline.

changing the sigma parameter from 0.05 to 0.3 with the degree parameter d = 1 (Figure S2), 2, and 3 (data not shown). Using the PC1 contribution rate from KPCA, we determined the sigma value (0.135) with the parameter d = 1 as the KPCA parameters for further analyses. We used these determined parameters to demonstrate that KPCA and conventional PCA yielded different profiles (Fig. 2). With conventional PCA, many samples were concentrated at particular positions on the score plot; therefore, it was difficult to identify any characteristics among samples. In contrast, with KPCA, the samples were holistically dispersed over the score plot, and the profiles tended to cluster according to individual differences. In this study, dispersion on the scores plot is very important to avoid biased grouping and generation of an unbalanced dataset. Thus, KPCA compared to PCA was suitable to use in subsequent analyses.

Important variable identifications of KPCA by incorporation of cforest. In this study, KPCA was used for unsupervised clustering (grouping) of the dataset with no class information in a data-driven manner. The grouping is key for this analytical procedure, but KPCA cannot calculate importance of variables directly due to an inner product computation process. To overcome this limitation, we used cforest to determine the key metabolites according to the importance in a model constructed by machine learning (random forest)[34]. cforest is an unbiased tree algorithm that overcomes a major limitation of the classical random forest approach involving variable selection bias[32]. To incorporate cforest into KPCA, profiles on the KPCA score plot were mathematically classified according to principal component (PC) plus and minus signs. In the present study, four groups based on the signs of PC1 and PC2 were manually generated for the cforest analysis (Fig. 3A). In this grouping, all samples were categorized in one of the 4 classes considered in the calculation. The number of samples belonging to classes 1, 2, 3, and 4 were 73, 94, 96, and 123, respectively. The samples categorized to classes 1 and 2 had a tendency to consume vegetable and fruit diets in the previous day, whereas the samples categorized to classes 3 and 4 had a tendency to consume protein- and fat-rich diets in the previous day (Table S1). Among this tendency, the class 1

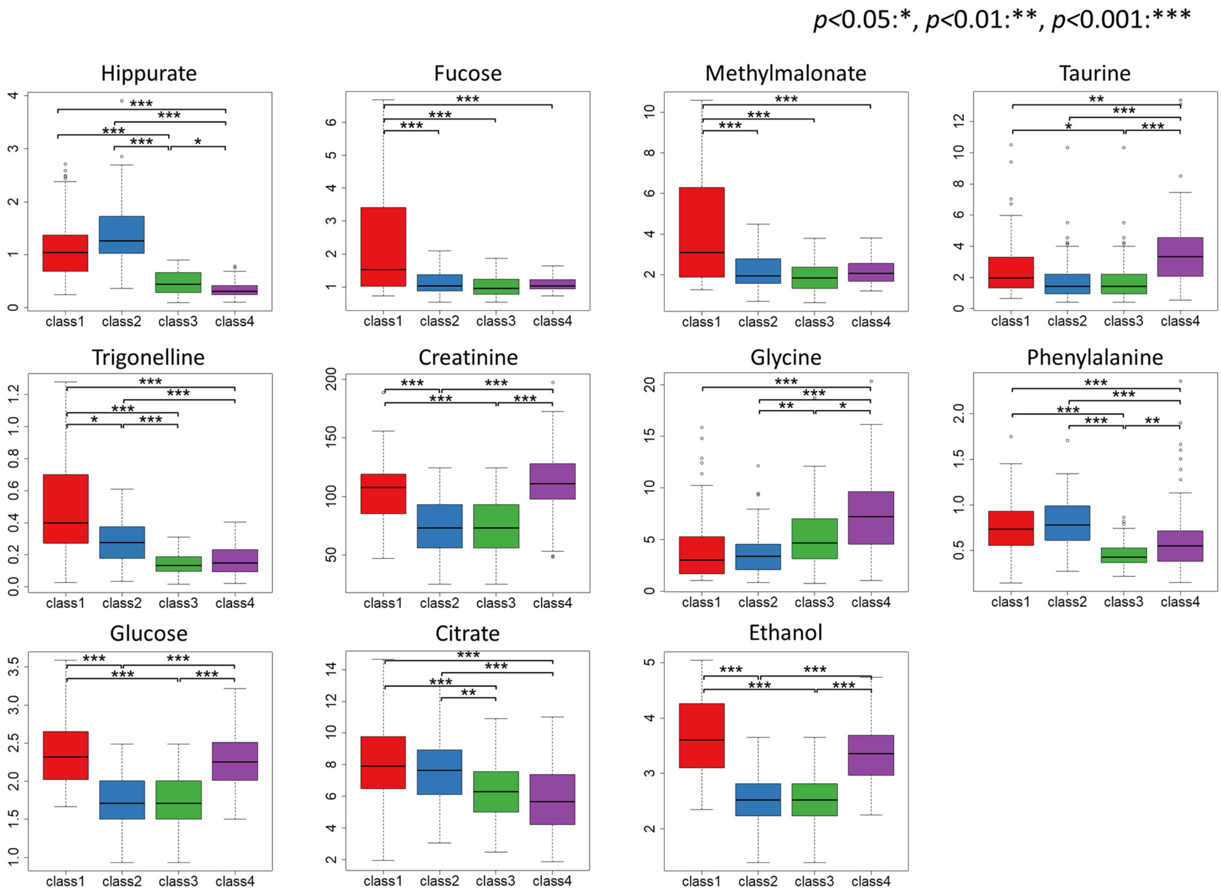

Figure 4. Box plots of the peak intensities of important variables. Significance: $p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$.

was likely to be also influenced by the alcohol intakes, and the class 3 was likely to be associated with fish intakes in the previous day. From the unsupervised grouping by KPCA in a data-driven manner, the class information was added to the original data for calculation of cforest modeling. The cforest modeling was performed with leave-one-out cross validation, resulting in 85.8% classification accuracy based on 4 classes with the confusion matrix shown in Table S2. The importance was also calculated for all variables in the cforest analysis, and the variables were aligned in descending order according to importance; only the top 100 variables are depicted in Fig. 3B. Among metabolites and inorganic elements, hippurate was identified as the most important variable, followed by fucose, methylmalonate, taurine, trigonelline, creatinine, glycine, glucose, citrate, phenylalanine, and ethanol. Significant intergroup differences were calculated to validate important variables (metabolites) (Fig. 4). For example, hippurate was significantly abundant in samples located in the positive PC2 group on the KPCA score plot compared with those in the PC2 negative group. In our previous study[35], hippurate was detected in NMR spectra but was not focused because there was no correlation with nutritional trends (high and low protein diets) studied in the previous paper. Thus, this current study enabled to provide a different perspective (a standpoint based on hippurate) that couldn't detect the relationship by the conventional method in the previous paper.

Evaluation of hippurate in input–output responses. Hippurate was identified as the most important variable contributing to KPCA class information. To characterize biological contents, hippurate, other urinary metabolites and inorganic elements, and nutritional data derived from daily dietary intake records were subjected to an MBA. MBA enables to screen direct or indirect relationships though MBA manages occasionally to detect any biological irrelevant and meaningless correlations. Notably, hippurate was associated with high concentrations of some vitamins and minerals present in foods eaten during the previous day (Figure S3A). Hippurate is abundant in diets that gain a lot of nourishment from foods containing aromatic compounds (e.g., polyphenols, aromatic side-chain amino acids) such as plants[36,37]. Studies in germ-free mice, which excrete low levels of hippurate in urine, suggest that aromatic compounds from dietary components are metabolized by symbiotic gut microbiota[38]. Two other reports have described a relationship between high levels of urinary hippurate excretion and intake of certain foods including fruits, vegetables, and whole-grain wheat flour[39,40]. In our study, the intake of fruits, vegetables, and whole-grain wheat flour during the previous day was considered to indicate a high intake of nutritive components (e.g., vitamins, minerals, food fiber, and carbohydrates), and thus our observations were consistent with those of previous reports[39,40]. Additionally, fruits such as banana and citrus fruit contain aromatic compounds[41], and hippurate excretion was found to increase over time after the consumption of orange juice[42],

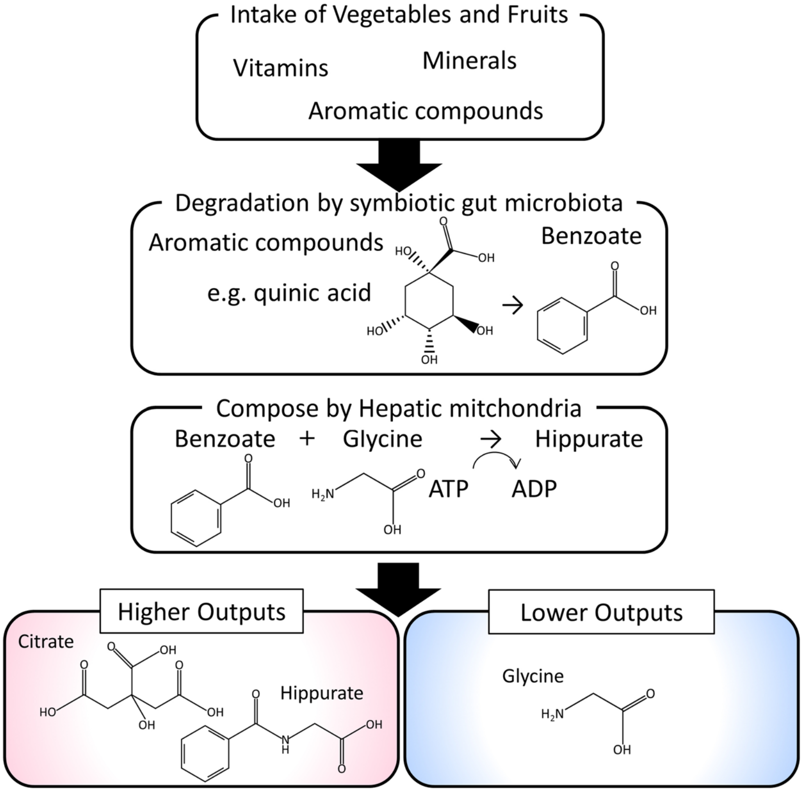

Figure 5. Putative pathway from dietary intake to metabolite generation. Significant metabolites were extracted using kernel principal component analysis followed by cforest, and associated intake nutrients were computed using market basket analysis.

a finding that was also consistent with our observation of high levels of urinary hippurate excretion following intake of edible fruits during the previous day. Overall, our data suggest that high hippurate levels observed in this study were consequent to the intake of a wide variety of vegetables and fruits.

An increasing hippurate level was also associated with increasing or decreasing levels of some output metabolites and minerals (Figure S3B). Among these metabolites and minerals, increasing levels of phenylalanine, tryptophan, and citrate have been reported to reflect the intake of whole-grain wheat flour[39]. Increasing levels of tryptophan and phenylalanine are attributed to the gut microbiotic shikimate pathway[43,44] which resynthesize aromatic amino acids[45].

One report suggests a relationship between reduced creatinine levels and vegetable intake[40] as well as between hippurate production and increasing and decreasing levels of citrate and glycine, respectively. The latter is particularly relevant, as hippurate is produced from benzoate and glycine in the human liver[46]. In addition, benzoate is derived from gut microbial degradation of aromatic compounds from vegetables or fruits[37], and citrate is an intermediate component of the TCA cycle. Hippurate production requires ATP; the increased citrate production resulting from an increased demand for ATP may explain the observed concurrent increases in hippurate and citrate levels[47]. Accordingly, the observed associations of hippurate with food intake and other metabolites were summarized into a putative simple pathway (Fig. 5). Overall, the analytical approach described here enabled us to capture input–output responses that were undetectable using linear PCA in previous studies[33,35].

In this study, only two principle components (PC1 and PC2) were used for the categorization of KPCA in the analytical procedure because it is important to evaluate with as few components as possible in terms of dimensional (data) reduction. However, this may be not always suitable in some cases to obtain a best performance of the analytical procedure. Moreover, the grouping to 4 classes may be not always better in some cases although characteristic features in the dataset were able to be captured by the 4 class grouping with appropriate dispersion on the scores plot in this study. Therefore, it will be beneficial and effective to develop a method for automatic determination of optimal number of components and classes and to incorporate the method into the analytical procedure in the future.

To evaluate the generality and robustness of our analytical approach, a dataset obtained from skin microbiota profiling[48] as another kind of omics data was used for the performance test. The same strategy was applied to this dataset, resulting in class information obtained from the KPCA scores plot (Figure S4A). In this categorization, the samples categorized in class 1 and in class 4 were mostly derived from the arm and from the face and neck, respectively, whereas those categorized in class 2 and in class 3 were mainly derived from the armpit, buttock, and leg (Figure S4B). The important variables contributing to each class were calculated by cforest analysis (Figure S4C), resulting in several key bacteria such as Corynebacteriaceae and Neisseriaceae identified as having high importance for the KPCA categorization. The important bacteria were further assessed by significant test among each class (Figure S5), revealing that the bacteria belonging to Corynebacteriaceae and Tissierellaceae were abundant on the arm (class 1) whereas the bacteria belonging to Propionibacteriaceae, Streptococcaceae, and Pasteurellaceae were abundantly located on the face and neck (class 4). The bacteria belonging to Moraxellaceae were relatively abundant on the armpit, buttock, and leg (classes 3 and 4) compared to the arm, face, and neck. From this analysis, our analytical approach enabled us to capture the localizations of bacteria on the body that were undetectable using linear PCA in a previous study[48]. Therefore, this approach should be useful not only for metabolic and microbiota profiling but also for profiling in other areas such as proteomics and transcriptomics in biological systems and environmental ecosystems.

Conclusions

This study established an analytical approach based on the combined use of non-linear KPCA and cforest with validation of the extracted important variables and subsequent evaluation of detected metabolites performed by MBA to identify input–output responses in humans. This approach enabled the identification of relationships between dietary intake and metabolites that could not be detected using linear PCA. By changing the kernel functions and parameters, this novel analytical approach could potentially be applied to a wide range of analyses in which useful and valuable information is extracted from biological and environmental systems. This approach, which can be applied to non-linear trend data, should therefore be incorporated as a new analytical option in diverse fields of science (especially life sciences).

Methods

Data preparations. In this study, we used 386 NMR and 386 ICP-OES datasets of urine samples collected from 8 human volunteers and 309 nutritional datasets of daily dietary intake records obtained from previous studies[33,35], and also spectral data acquired in the present study. The human ethical committees of RIKEN Yokohama Research Institute and Yokohama City University approved this study which enrolled human subjects who provided informed consent. All methods and procedures were performed in accordance with the relevant guidelines and regulations.

Data processing. Collected ^{1}H NMR data (32,000 data points) were normalized via probabilistic quotient normalization[49] using the mQTL package (Revolution R open software, 8.0.1 beta 64-bit) and aligned using the *i*coshift[50] program on Matlab R2015b (MathWorks Japan, Tokyo, Japan) in an in-house computing environment. Peak-picked NMR data and ICP-OES data were merged into a data matrix with auto scaling for further analyses.

KPCA. As mentioned above, KPCA was developed for non-linear PCA[28]. Accordingly, KPCA comprises PCA, in which a non-linear kernel function has been incorporated, allowing the performance of non-linear PCA using matrices converted from input matrices by the kernel function. During processing, a kernel-enabled non-linear data approximation enables the extraction of information that differs from that obtained using conventional linear approximations. Although several methods have been developed for the kernel function, ANOVA kernel method was the best performance of dispersion on the scores plot compared to the other kernel methods, i.e., Gaussian (Figure S6), Laplace (Figure S7), and Bessel (Figure S8). Thus, the following ANOVA kernel method was used in the present study:

$$\text{ANOVA kernel: } K(x,\, y) = \left(\sum\nolimits_{(k=1)}^{n} exp(-\sigma(x^k - x'^k)^2)\right)^d \tag{1}$$

ANOVA kernel calculations were performed using the kpca function installed in the R kernlab package[51].

cforest. The random forest method[34] is a well-known machine learning algorithm for clustering and regression analyses and has become widely used in recent years in bioinformatics studies. Random forest can be applied to datasets with non-linear features and is intended for the construction of predictive and discriminant models as well as the calculation of important variables for constructing predictors. However, random forest is associated with the potential for bias caused by differences in sample numbers between groups. In other words, a group with a larger number of samples is more likely to be identified as having greater importance with respect to corresponding variables[32,52,53]. To counteract this bias, the cforest algorithm was developed based on a non-biased decision tree which overcomes the weakness associated with variable selections using conditional inference trees to calculate a permutation importance. In this study, cforest was calculated using the cforest function in the R party software package[54] for original data with group information determined from KPCA results. In this process, we used the tuneRF function in the R randomForest package to tune ntree (number of decision trees) set to 80 and mtry (number of features used to make a decision tree) set to 900. Leave-one-out cross validation was performed for verification of the constructed model.

Statistical analysis. MBA with NMR, ICP-OES, and nutrient variables was performed using the R software package arules[55] as previously described[33]. The association rules were determined to exceed the cut-off values of 0.0625 for support, 0.25 for confidence, and 1.2 for lift. The association network was drawn using the Cytoscape program[56]. The Kruskal–Wallis test was used to determine significant differences between groups.

Analytical protocol. For analysis of KPCA, the R "kernlab" package is installed from the CRAN website. Then the command "library(kernlab)" was executed for loading in the R platform. The function "kpca" was executed with kernel="anovadot" for ANOVA function. The obtained KPCA scores were manually classified into four groups based on PC1 and PC2 plus and minus signs. Class information for four groups was added to the original data, followed by execution of the cforest program using the R "party" package installed from the CRAN website. The function "cforest" was executed, and variable importances were calculated. Finally, the variables were sorted in descending order according to their importances for further analyses such as significance tests and MBA. The R protocols (i.e., KPCA, cforest, and MBA) used in this study were deposited on our website (http://dmar.riken.jp/Rscripts/).

References

1. Wong, A. *et al*. muHigh resolution-magic-angle spinning NMR spectroscopy for metabolic phenotyping of Caenorhabditis elegans. *Anal Chem* **86**, 6064–6070, https://doi.org/10.1021/ac501208z (2014).
2. Guennec, A. L., Giraudeau, P. & Caldarelli, S. Evaluation of fast 2D NMR for metabolomics. *Anal Chem* **86**, 5946–5954, https://doi.org/10.1021/ac500966e (2014).
3. Larive, C. K., Barding, G. A. Jr & Dinges, M. M. NMR spectroscopy for metabolomics and metabolic profiling. *Anal Chem* **87**, 133–146, https://doi.org/10.1021/ac504075g (2015).
4. Dumas, M. E. *et al*. Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study. *Anal Chem* **78**, 2199–2208, https://doi.org/10.1021/ac0517085 (2006).
5. Viant, M. R. *et al*. International NMR-based environmental metabolomics intercomparison exercise. *Environ Sci Technol* **43**, 219–225 (2009).
6. Ogawa, D. M. *et al*. Biogeochemical typing of paddy field by a data-driven approach revealing sub-systems within a complex environment–a pipeline to filtrate, organize and frame massive dataset from multi-omics analyses. *PLoS One* **9**, e110723, https://doi.org/10.1371/journal.pone.0110723 (2014).
7. Yoshida, S., Date, Y., Akama, M. & Kikuchi, J. Comparative metabolomic and ionomic approach for abundant fishes in estuarine environments of Japan. *Sci Rep* **4**, 7005, https://doi.org/10.1038/srep07005 (2014).
8. Asakura, T., Date, Y. & Kikuchi, J. Comparative analysis of chemical and microbial profiles in estuarine sediments sampled from Kanto and Tohoku regions in Japan. *Anal Chem* **86**, 5425–5432, https://doi.org/10.1021/ac5005037 (2014).
9. Asakura, T., Sakata, K., Yoshida, S., Date, Y. & Kikuchi, J. Noninvasive analysis of metabolic changes following nutrient input into diverse fish species, as investigated by metabolic and microbial profiling approaches. *PeerJ* **2**, e550, https://doi.org/10.7717/peerj.550 (2014).
10. Ito, K., Sakata, K., Date, Y. & Kikuchi, J. Integrated analysis of seaweed components during seasonal fluctuation by data mining across heterogeneous chemical measurements with network visualization. *Anal Chem* **86**, 1098–1105, https://doi.org/10.1021/ac402869b (2014).
11. Wei, F., Ito, K., Sakata, K., Date, Y. & Kikuchi, J. Pretreatment and integrated analysis of spectral data reveal seaweed similarities based on chemical diversity. *Anal Chem* **87**, 2819–2826, https://doi.org/10.1021/ac504211n (2015).
12. Date, Y., Iikura, T., Yamazawa, A., Moriya, S. & Kikuchi, J. Metabolic Sequences of Anaerobic Fermentation on Glucose-Based Feeding Substrates Based on Correlation Analyses of Microbial and Metabolite Profiling. *J Proteome Res* **11**, 5602–5610, https://doi.org/10.1021/Pr3008682 (2012).
13. Ogura, T., Date, Y., Tsuboi, Y. & Kikuchi, J. Metabolic dynamics analysis by massive data integration: application to tsunami-affected field soils in Japan. *ACS Chem Biol* **10**, 1908–1915, https://doi.org/10.1021/cb500609p (2015).
14. Wishart, D. S. *et al*. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res* **41**, D801–807, https://doi.org/10.1093/nar/gks1065 (2013).
15. Cui, Q. *et al*. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat Biotechnol* **26**, 162–164, https://doi.org/10.1038/nbt0208-162 (2008).
16. Bingol, K., Zhang, F., Bruschweiler-Li, L. & Bruschweiler, R. TOCCATA: a customized carbon total correlation spectroscopy NMR metabolomics database. *Anal Chem* **84**, 9395–9401, https://doi.org/10.1021/ac302197e (2012).
17. Bingol, K., Zhang, F., Bruschweiler-Li, L. & Bruschweiler, R. Quantitative analysis of metabolic mixtures by two-dimensional 13C constant-time TOCSY NMR spectroscopy. *Anal Chem* **85**, 6414–6420, https://doi.org/10.1021/ac400913m (2013).
18. Chikayama, E. *et al*. Statistical indices for simultaneous large-scale metabolite detections for a single NMR spectrum. *Anal Chem* **82**, 1653–1658, https://doi.org/10.1021/ac9022023 (2010).
19. Kikuchi, J. *et al*. SpinCouple: Development of a Web Tool for Analyzing Metabolite Mixtures via Two-Dimensional J-Resolved NMR Database. *Anal Chem* **88**, 659–665, https://doi.org/10.1021/acs.analchem.5b02311 (2016).
20. Xia, J., Sinelnikov, I. V., Han, B. & Wishart, D. S. MetaboAnalyst 3.0–making metabolomics more meaningful. *Nucleic Acids Res* **43**, W251–257, https://doi.org/10.1093/nar/gkv380 (2015).
21. Steinbeck, C. & Kuhn, S. NMRShiftDB–compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **65**, 2711–2717, https://doi.org/10.1016/j.phytochem.2004.08.027 (2004).
22. Worley, B. & Powers, R. MVAPACK: a complete data handling package for NMR metabolomics. *ACS Chem Biol* **9**, 1138–1144, https://doi.org/10.1021/cb4008937 (2014).
23. Lewis, I. A., Schommer, S. C. & Markley, J. L. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn Reson Chem* **47**(Suppl 1), S123–126, https://doi.org/10.1002/mrc.2526 (2009).
24. Hao, J., Astle, W., De Iorio, M. & Ebbels, T. M. BATMAN–an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* **28**, 2088–2090, https://doi.org/10.1093/bioinformatics/bts308 (2012).
25. Cloarec, O. *et al*. Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic 1H NMR Data Sets. *Anal Chem* **77**, 1282–1289, https://doi.org/10.1021/ac048630x (2005).
26. Ito, K., Tsutsumi, Y., Date, Y. & Kikuchi, J. Fragment Assembly Approach Based on Graph/Network Theory with Quantum Chemistry Verifications for Assigning Multidimensional NMR Signals in Metabolite Mixtures. *ACS Chem Biol* **11**, 1030–1038, https://doi.org/10.1021/acschembio.5b00894 (2016).
27. Misawa, T., Komatsu, T., Date, Y. & Kikuchi, J. SENSI: signal enhancement by spectral integration for the analysis of metabolic mixtures. *Chem Commun (Camb)* **52**, 2964–2967, https://doi.org/10.1039/c5cc09442a (2016).
28. Scholkopf, B., Smola, A. & Muller, K. R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**, 1299–1319, https://doi.org/10.1162/089976698300017467 (1998).
29. Cho, H. W. *et al*. Discovery of metabolite features for the modelling and analysis of high-resolution NMR spectra. *Int J Data Min Bioinform* **2**, 176–192 (2008).
30. Ginsburg, S., Ali, S., Lee, G., Basavanhally, A. & Madabhushi, A. Variable importance in nonlinear kernels (VINK): classification of digitized histopathology. *Med Image Comput Comput Assist Interv* **16**, 238–245 (2013).
31. Rakotomamonjy, R. Variable Selection Using SVM-based Criteria. *Journal of Machine Learning Research* **3**, 1357–1370 (2003).
32. Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics* **8**, 25, https://doi.org/10.1186/1471-2105-8-25 (2007).
33. Shiokawa, Y., Misawa, T., Date, Y. & Kikuchi, J. Application of Market Basket Analysis for the Visualization of Transaction Data Based on Human Lifestyle and Spectroscopic Measurements. *Anal Chem* **88**, 2714–2719, https://doi.org/10.1021/acs.analchem.5b04182 (2016).
34. Breiman, L. Random Forest. *Machine Learning* **45**, 5–32 (2001).

35. Misawa, T., Date, Y. & Kikuchi, J. Human metabolic, mineral, and microbiota fluctuations across daily nutritional intake visualized by a data-driven approach. *J Proteome Res* **14**, 1526–1534, https://doi.org/10.1021/pr501194k (2015).
36. Hertog, M. G. L., Hollman, P. C. H., Katan, M. B. & Kromhout, D. Intake of potentially anticarcinogenic flavonoids and their determinants in adults in the Netherlands. *Nutrition and Cancer* **20**, 21–29, https://doi.org/10.1080/01635589309514267 (1993).
37. Lees, H. J., Swann, J. R., Wilson, I. D., Nicholson, J. K. & Holmes, E. Hippurate: the natural history of a mammalian-microbial cometabolite. *J Proteome Res* **12**, 1527–1546, https://doi.org/10.1021/pr300900b (2013).
38. Claus, S. P. *et al.* Systemic multicompartmental effects of the gut microbiome on mouse metabolic phenotypes. *Mol Syst Biol* **4**, 219, https://doi.org/10.1038/msb.2008.56 (2008).
39. Fardet, A. *et al.* Whole-grain and refined wheat flours show distinct metabolic profiles in rats as assessed by a 1H NMR-based metabonomic approach. *J Nutr* **137**, 923–929 (2007).
40. Walsh, M. C. *et al.* Influence of acute phytochemical intake on human urinary metabolomic profiles. *Am J Clin Nutr* **86**, 1687–1693 (2007).
41. Kumar, N. & Pruthi, V. Potential applications of ferulic acid from natural sources. *Biotechnology Reports* **4**, 86–93, https://doi.org/10.1016/j.btre.2014.09.002 (2014).
42. Pereira-Caro, G. *et al.* Orange juice (poly)phenols are highly bioavailable in humans. *Am J Clin Nutr* **100**, 1378–1384, https://doi.org/10.3945/ajcn.114.090282 (2014).
43. Herrmann, K. M. The Shikimate Pathway: Early Steps in the Biosynthesis of Aromatic Compounds. *The Plant Cell* **7**, 907–919 (1995).
44. Herrmann, K. M. & Weaver, L. M. The Shikimate Pathway. *Annual review of plant physiology and plant molecular biology* **50**, 473–503, https://doi.org/10.1146/annurev.arplant.50.1.473 (1999).
45. Fuller, M. F. & Reeds, P. J. Nitrogen cycling in the gut. *Annual review of nutrition* **18**, 385–411, https://doi.org/10.1146/annurev.nutr.18.1.385 (1998).
46. Gatley, S. J. & Sherratt, H. S. The synthesis of hippurate from benzoate and glycine by rat liver mitochondria. Submitochondrial localization and kinetics. *The Biochemical journal* **166**, 39–47 (1977).
47. Krahenbuhl, L., Reichen, J., Talos, C. & Krahenbuhl, S. Benzoic acid metabolism reflects hepatic mitochondrial function in rats with long-term extrahepatic cholestasis. *Hepatology* **25**, 278–283, https://doi.org/10.1053/jhep.1997.v25.pm0009021934 (1997).
48. Tsutsui, S., Date, Y. & Kikuchi, J. Visualizing Individual and Region-specific Microbial–metabolite Relations by Important Variable Selection Using Machine Learning Approaches. *Journal of Computer Aided Chemistry* **18**, 31–41 (2017).
49. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* **78**, 4281–4290, https://doi.org/10.1021/ac051632c (2006).
50. Savorani, F., Tomasi, G. & Engelsen, S. B. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J Magn Reson* **202**, 190–202, https://doi.org/10.1016/j.jmr.2009.11.012 (2010).
51. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab-An S4 Package for Kernel Methods in R. *Journal of Staistical Software* **11**, 1–20 (2004).
52. Hothorn, T., Hornik, K. & Zeileis, A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* **15**, 651–674 (2006).
53. Kim, H. & Loh, W.-Y. Classification Trees With Unbiased Multiway Splits. *J. Amer. Statist. Assoc* **96**, 598–604 (2001).
54. Hothorn, T., Hornik, K. & Zeileis, A. party: A Laboratory for Recursive Partytioning. http://CRAN.R-project.org/, [R package version0.9–0] (2006).
55. Michael, H., Sundheer, C., Kurt, H. & Christian, B. The arules R-Package ecosystem: analyzing interesting patterns from large transaction data sets. *Journal of Machine Learning Research* **12**, 2021–2025 (2011).
56. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504, https://doi.org/10.1101/gr.1239303 (2003).

Acknowledgements

The authors wish to thank Takuma Misawa (Yokohama City University) for supplying urinary and nutrient data. This research was supported in part by a Grant-in-Aid for Scientific Research to J.K.

Author Contributions

Y.S. performed the experiments, analyzed the data, and wrote the paper. Y.D. assisted the data analysis and wrote the paper. J.K. supervised this study.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-20121-w.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.