Supervised Multivariate and Univariate Analyses for NMR-based Metabolic Profiling to Explore Characteristic Metabolites among Sample Groups

Product used: Nuclear Magnetic Resonance (NMR)

Recently, NMR metabolic profiling (NMR-MP) techniques, which are useful for systematically analyzing metabolites in an organism (metabolome), have been widely used in post-genome studies and quality controls for agricultural and biotechnological products. NMR-MP is characterized as a non-target metabolome analysis. It is difficult to interpret results directly from data in non-target analysis; therefore, multivariate analyses are applied to explore characteristic information from multivariate data. This is referred to as “data-mining”.

It is important to choose appropriate multivariate analyses that correspond to the goal of your study. In this note (O)PLS-DA multivariate analysis using sample group as an objective variable and univariate analyses with NMR metabolome data were employed to explore characteristic metabolites among sample groups (i.e. marker molecules) and build a discrimination model based on metabolites.

Materials and Methods

Broccoli sprouts growing in two different environments (w/ light or w/o light) were used as an example for multivariate with two different sample groups. Broccoli seeds were sowed on 1 wt% ager gel. Until 5 days after seeding, they were cultivated without light. From 5 days until 10 days after seeding, half of them were cultivated with light and another half were cultivated without light. Then seedlings were harvested on 10 days after seeding. Polar metabolites in two different sprouts were extracted by a modified Bligh-Dyer method. The extracts were then dried with a centrifugal evaporator. Dried metabolite mixtures were resuspended with the NMR solvent (a 100 mM deuterated potassium phosphate buffer, pH = 7.0). ^{1}H single pulse spectra were recorded and used for further analyses (**Figure 1**). Eight sample replicates in each sample group were performed.

Bucket integrations were performed on the series of spectra (0.04 ppm width each, between 0.5-9.0 ppm). As a result, the matrix (multivariate for multivariate analyses) consisted of 194 variables * 18 samples were obtained. OPLS-DA and univariate analyses were performed with the “ropls” package in bioconductor (https://www.bioconductor.org/) [1] and “muma” package (https://CRAN.R-project.org/package=muma) [2] in the “R” which is a statistical scripting language.

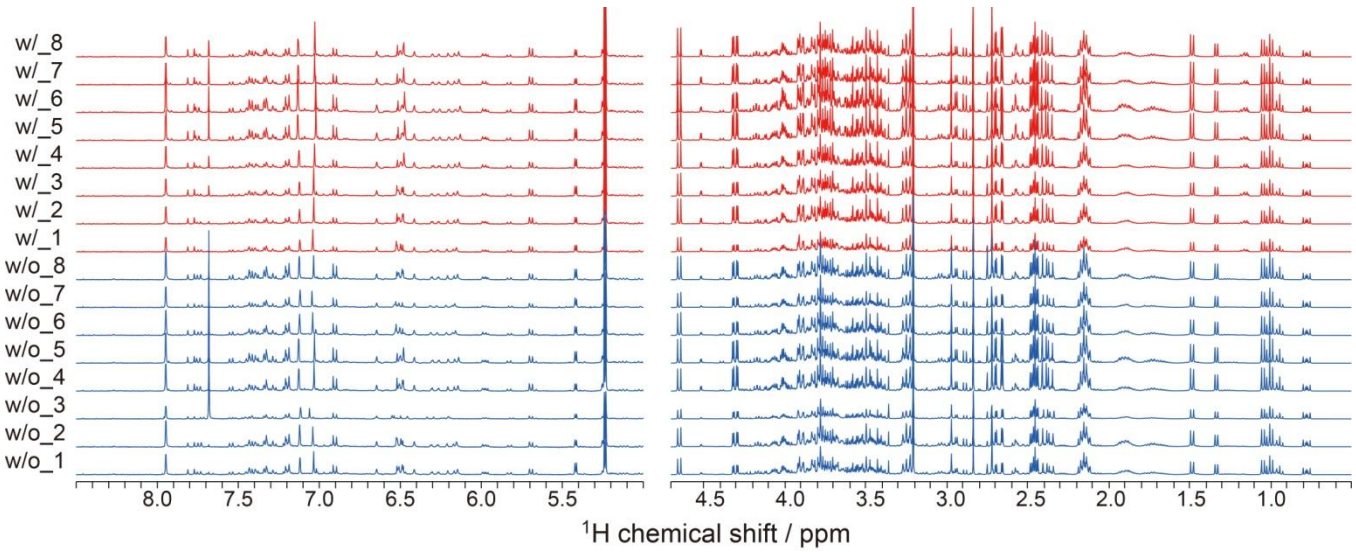

Figure 1. Series of ^{1}H spectra of polar metabolites in broccoli sprouts cultivated w/ or w/o light. Y range in aromatic region (8.5-5.0 ppm) was scaled with 8 time larger than that in aliphatic region (4.7-0.5 ppm). Eight experimental replicates were performed in each sample group. NMR spectra were recorded continuously on a JNM-ECZ400S spectrometer equipped with a ROYAL probe and ASC30 auto sample changer.

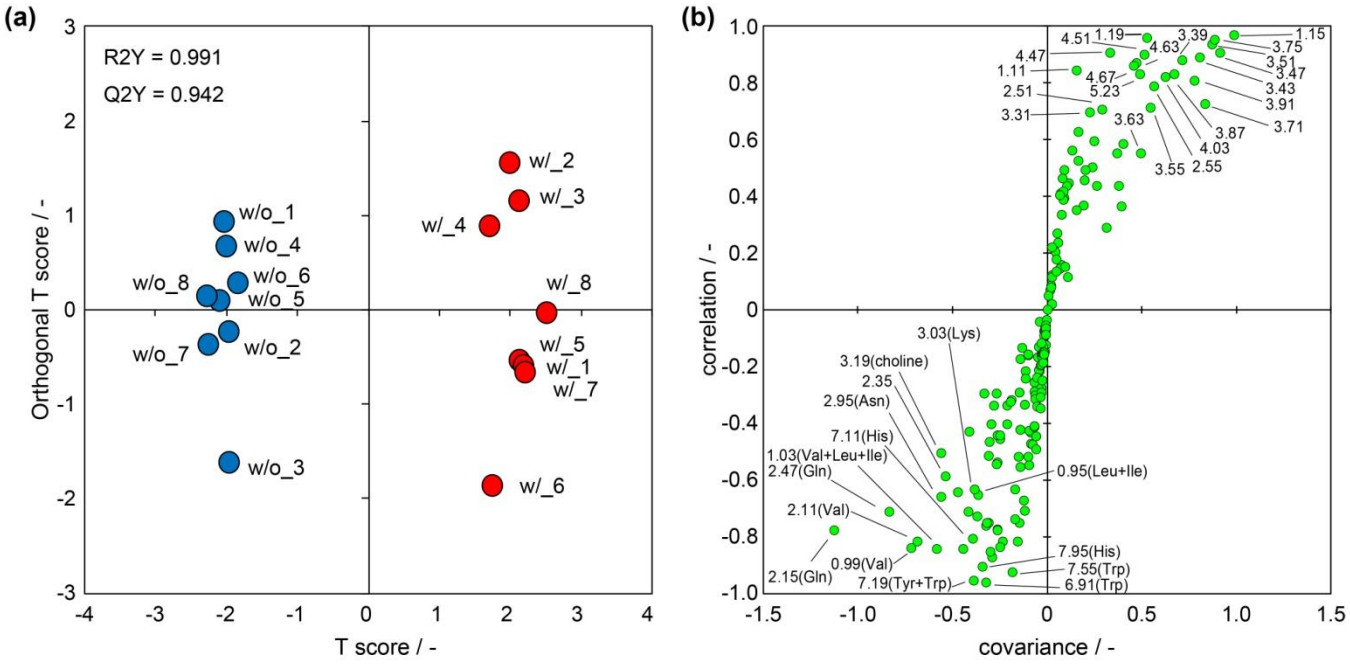

Figure 2. OPLS-DA using ^{1}H NMR spectra of polar metabolites in sprouts growing two different conditions. (a) score plot. (b) S-plot. Normalizing with sum of all variables and Pareto scaling with centering were performed before OPLS-DA.

Applying Discrimination Analysis

In most cases in NMR-MP, the number of variables is far larger than that of samples; that is why the linear discrimination analysis is not suitable for NMR-MP. In that cases, PLS-DA is effective which is a discrimination analysis using Partial Least Squares model [3]. In linear discrimination analysis, a discrimination function is calculated as linear combinations which maximize a variation among sample groups. That is similar to procedures in Principal Component Analysis (PCA) as shown in previous our application note [4] while PCA maximize a variance of principal components. However maximizing a variance doesn't means maximizing variation among sample groups. In PLS-DA, linear combinations which maximize covariance between data and objective variable corresponding to the sample group are selected. In turn, the resultant discrimination functions in PLS-DA maximize the variations among the sample group.

In this note, OPLS-DA [5], which provides results which are easier to interpret than PLS-DA, was employed (**Figure 2**). Each sample was normalized using sum of all variables and centering and Pareto scaling were performed on each variable.

In the score plot, sprouts growing w/ light and w/o were clearly discriminated in first T score (**Figure 2(a)**). A explained variation (R^2Y) and predicted variation (Q^2Y) were 0.991 and 0.942, respectively. That indicates resultant OPLS model was valid.

S-plot is a scatter plot using covariance (x-axis) and correlation (y-axis) between scores and variables. Therefore metabolites high amount in sprout w/ light are in the first quadrant, and those high in w/o light are in the third quadrant. In the first quadrant, there were natural sugars including glucose and sucrose. In the third quadrant, there were amino acid including glutamine, asparagine, valine. That would reflect autotrophic metabolism in sprout w/ light, meanwhile, in sprout w/o light, energy metabolism depended heavily on providing amino acid from degradation of protein. Note that NMR-MP doesn't provide any evidence of facts (ex. we are not sure high amino acids in sprout w/o light were actually caused by degradation of proteins). However NMR-MP, data-driven approach, provides some hypothesis (i.e. high amino acids in sprout w/o light would be caused by degradation of proteins for energy recycling). You have to perform verification experiments to confirm the facts. Nonetheless, data-driven approaches such as NMR-MP provide opportunities to find out a new fact which we cannot fined out using a conventional hypothesis-driven approach.

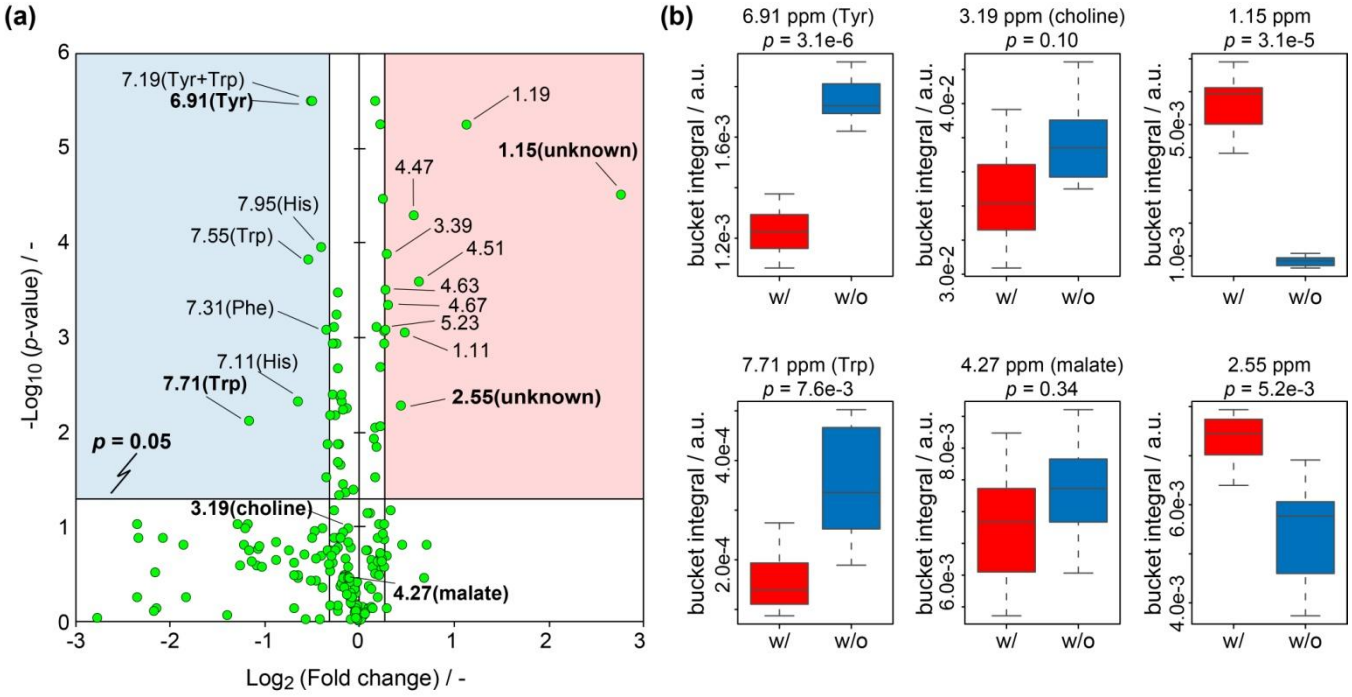

Figure 3. Univariate analysis using ^{1}H NMR spectra of polar metabolites in sprouts growing two different conditions. (a) volcano plot. The region highlighted with red ($p < 0.05$ and fold change < 0.833) indicates metabolites that are present in significant higher amounts in sprout w/ light, meanwhile the region highlighted with blue ($p < 0.05$ and fold change > 1.2) indicates metabolites here are significant higher in sprout w/o light. (b) box plots. Box plots visualize distribution of each variable in each sample.

Applying Univariate Analyses

Statistical hypothesis testing for a difference of averages in particular variable between two sample group is typical analysis procedure in modern laboratories. This is referred to as "univariate analysis" contrasted with multivariate analysis. Here the univariate analysis were performed for all variables repeatedly (i.e. multiple comparison). We employed a "Volcano plot" which can visualized multiple univariate analyses comprehensively.

Before hypothesis testing for a difference of averages, Shapiro Wilk's test was performed to make sure whether each variable was according normal distribution or not. The Welch's T test was performed on variables according normal distribution, meanwhile the Mann-Whitney's U test was performed on variables not according normal distribution in the Shapiro Wilk's test. *P*-values in two hypothesis test for a difference of averages were adjusted with the Benjamini-Hochberg method (= adjusting false discovery rate) to avoid increasing type I errors by multiple comparison. The adjusted *p*-values were used in the volcano plot. Methodological details for identify variables having significant difference among sample groups are described in the reference material [6].

While the multiple comparison provides a variable which is significantly higher in particular sample group in each, it is not easy to overview whole results in all variables together. The volcano plot enable that. The volcano plot is a scatter plot in which x- and y-axis represents fold change and *p*-value of each variable, respectively. Fold changes and *p*-values are typically shown as log2 and –log10, respectively, for making plots wide spread. As a result, the scatter plot looks like "volcano" (**Figure 3(a)**). "muma" package in R automatically performs several statistical hypothesis testing, adjusting *p*-values, making a volcano plot.

The volcano plot from ^{1}H NMR spectra of polar metabolites in sprouts growing two different conditions are shown in **Figure 3(a)**. Fold changes of 0.8 and 1.2 and *p*-value of 0.05% were used as boundaries. Therefore the region highlighted with red in **Figure 3(a)** ($p < 0.05$ and fold change > 1.2) indicates metabolites here are significant higher in sprout w/ light, meanwhile the region highlighted with blue ($p < 0.05$ and fold change < 0.8) indicates metabolites here are significant higher in sprout w/o light. Signals at 1.15, 1.19, 4.51 ppm (unfortunately they were not identified into particular metabolite) were more in sprouts growing w/ light, and signals at 7.19, 6.19, 7.95, 7.55, 7.31 ppm (they were aromatic amino acids including His, Tyr, Phe, Trp) were more in sprouts growing w/o light,

We can see distributions of amount in each variable by use of box plots (**Figure 3(b)**). Box plots indicate minimum, 25 percentile, 50 percentile, 75 percentile, and maximum of a variable in each sample group. Here a range between 25 and 75 percentile is shown as box and a range between minimum and 25 percentile and a range between 75 percentile and maximum are shown as whiskers. For examples box plots in 6.91 and 7.71 ppm shown in upper and lower left (Tyr and Trp) indicate distributions in 8 replicates of bucket integral in plants without light are higher that that with light. Meanwhile box plots in 1.15 and 2.55 ppm shown in upper and lower right indicate distributions in plants with light are higher than that without light. Also box plots in 3.19 and 4.27 ppm shown in upper and lower middle (choline and malate) indicate distributions in plants withot light are slightly higher than that without light but distributions with light are basically overlapped with that without light .

Multivariate analysis using sample categories as an objective valuable, as shown in current note, is fruitful to explore maker molecules or built discrimination model. Also univariate analyses with volcano plot is fruitful even for multivariate data. We can perform the same analyses with data with more than three sample groups. Meanwhile regression analyses should be performed with data in which quantitative variables are used as objective variables.

References

[1] Gentleman; R.C. et al., *Genome Biol.* (2004) **5**, R80. Thévenot; E.A., Roux; A., Xu; Y., Ezan; E., Junot; C., *J. Proteome Res.* (2015) **14**, 3322.
[2] Gaude; E., et al. *Curr. Metabolomics* (2013) **1**, 180.
[3] Barker; M. and Rayens; W*., J. Chemometrics* (2003) **17**, 166.
[4] Application note "*Unsupervised Multivariate Analyses for NMR-based Metabolomics*" (NM170013).
[5] Eriksson; L. et al., *J. Chemometrics* (2008) **22**, 594.
[6] Contento; A. L. et al., *Plant Physiol.* (2004) **135**, 2330.
[7] Goodpaster; A. M. et al., *Anal Biochem.* (2010) **401**, 134.

JEOL Ltd. 3-1-2 Musashino Akishima Tokyo 196-8558 Japan Sales Division Tel. +81-3-6262-3560 Fax. +81-3-6262-3577
www.jeol.com ISO 9001 • ISO 14001 Certified

• **AUSTRALIA & NEW ZEALAND** /JEOL(AUSTRALASIA) Pty.Ltd. Suite 1, L2 18 Aquatic Drive - Frenchs Forest NSW 2086 Australia • **BELGIUM** /JEOL (EUROPE) B.V. Planet II, Gebouw B Leuvensesteenweg 542, B-1930 Zaventem Belgium • **BRAZIL** /JEOL Brasil Instrumentos Cientificos Ltda. Av. Jabaquara, 2958 5º andar conjunto 52 ; 04046-500 Sao Paulo, SP Brazil • **CANADA** /JEOL CANADA, INC. 3275 1ere Rue, Local #8 St-Hubert, QC J3Y-8Y6, Canada • **CHINA** /JEOL(BEIJING) CO., LTD. Zhongkeziyuan Building South Tower 2F, Zhongguancun Nansanjie Street No. 6, Haidian District, Beijing, P.R.China • **EGYPT** /JEOL SERVICE BUREAU 3rd Fl. Nile Center Bldg., Nawal Street, Dokki, (Cairo), Egypt • **FRANCE** /JEOL (EUROPE) SAS Espace Claude Monet, 1 Allee de Giverny 78290, Croissy-sur-Seine, France • **GERMANY** /JEOL (GERMANY) GmbH Gute Aenger 30 85356 Freising, Germany • **GREAT BRITAIN & IRELAND** /JEOL (U.K.) LTD. JEOL House, Silver Court, Watchmead, Welwyn Garden City, Herts AL7 1LT, U.K. • **ITALY** /JEOL (ITALIA) S.p.A. Palazzo Pacinotti - Milano 3 City, Via Ludovico il Moro, 6/A 20080 Basiglio(MI) Italy • **KOREA** /JEOL KOREA LTD. Dongwoo Bldg. 7F, 1443, Yangjae Daero, Gangdong-Gu, Seoul, 05355, Korea • **MALAYSIA** /JEOL(MALAYSIA) SDN.BHD. 508, Block A, Level 5, Kelana Business Center, 97, Jalan SS 7/2, Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia • **MEXICO** /JEOL DE MEXICO S.A. DE C.V. Arkansas 11 Piso 2 Colonia Napoles Delegacion Benito Juarez, C.P. 03810 Mexico D.F., Mexico • **RUSSIA** /JEOL (RUS) LLC. Krasnoproletarskaya Street, 16, Bld. 2, 127473, Moscow, Russian Federation • **SCANDINAVIA** /SWEDEN JEOL (Nordic) AB Hammarbacken 6A, Box 716, 191 27 Sollentuna Sweden • **SINGAPORE** /JEOL ASIA PTE.LTD. 2 Corporation Road #01-12 Corporation Place Singapore 618494 • **TAIWAN** /JIE DONG CO., LTD. 7F, 112, Chung Hsiao East Road, Section 1, Taipei, Taiwan 10023 (R.O.C.) • **THE NETHERLANDS** /JEOL (EUROPE) B.V. Lireweg 4, NL-2153 PH Nieuw-Vennep, The Netherlands • **USA** /JEOL USA, INC. 11 Dearborn Road, Peabody, MA 01960, U.S.A.