Journal of Plant Biology, August 2007, 50(4) : 517-521

REPORT

Analysis of Metabolite Profile Data Using Batch-Learning Self-Organizing Maps

Jae Kwang Kim[1]*, Myoung Rae Cho[1], Hyung Jin Baek[1], Tae Hun Ryu[1], Chang Yeon Yu[2], Myong Jo Kim[2], Eiichiro Fukusaki[3], and Akio Kobayashi[3]

[1]National Institute of Agricultural Biotechnology, RDA, Suwon 441-707, Korea
[2]Bioherb Research Institute, Kangwon National University, Chunchon 200-701, Korea
[3]Department of Biotechnology, Graduate School of Engineering, Osaka University, Osaka 565-0871, Japan

Novel tools are needed for efficient analysis and visualization of the massive data sets associated with metabolomics. Here, we describe a batch-learning self-organizing map (BL-SOM) for metabolome informatics that makes the learning process and resulting map independent of the order of data input. This approach was successfully used in analyzing and organizing the metabolome data for *Arabidopsis thaliana* cells cultured under salt stress. Our 6 × 4 matrix presented patterns of metabolite levels at different time periods. A negative correlation was found between the levels of amino acids and metabolites related to glycolysis metabolism in response to this stress. Therefore, BL-SOM could be an excellent tool for clustering and visualizing high dimensional, complex metabolome data in a single map.

Keywords: batch-learning self-organizing map, cell culture, metabolome analysis, salt stress

Metabolomics is a comprehensive analysis during which thorough metabolite profiling is performed. This approach has emerged as a functional genomics methodology that contributes to our understanding of the complex molecular interactions in biological systems (Fiehn, 2002). Its final output is an exhaustive profile of all the metabolites present in a target organism. Unfortunately, this technique is still in an infant stage, and many of the necessary tools are not yet available (Sumner et al., 2003). The great diversity of chemical properties and the wide ranges of metabolite concentrations pose a significant challenge because these methods must be robust and reproducible so that samples can be reliably compared. In addition, metabolomics has not yet been widely implemented primarily because an efficient method is lacking for cataloging data into useful and functionally meaningful groups. Multivariate analysis with an appropriate algorithm could be performed depending on data structure and the mining intention. Principle component analysis (PCA) and hierarchical cluster analysis (HCA) are often conducted to systematically group related patterns of metabolite levels. Although these methods can effectively cluster genes with similar profiles, there is no direct relationship between different branches (Fiehn et al., 2000; Roessner et al., 2001; Sumner et al., 2003; Tikunov et al., 2005).

Kohonen's Self-Organizing Map (SOM) is an unsupervised neural network algorithm that has successfully been used to analyze very large data files in various fields, e.g., for process-monitoring and visualization, exploratory data analysis, and simulation of brain-like feature maps (Kohonen, 1982, 1990; Kohonen et al., 1996). However, the original SOM algorithm requires a long time for its calculation, and may produce different clustering results in its topology depending on the order of data input. This original SOM had now been improved as a batch-learning SOM (BL-SOM). The initial weight vectors are set by PCA and the learning process is designed to be independent of the order of input of vectors; hence, the result is reproducible (Kanaya et al., 2001; Abe et al., 2003). BL-SOM can be done in the laboratory using a personal computer because the algorithm does not require high CPU (central processing unit) power (Fukusaki and Kobayashi, 2005). Although it has been used in "omics" sciences, including genomics and transcriptomics (Kanaya et al., 2001; Abe et al., 2003, 2006; Hirai et al., 2004), only a few examples for BL-SOM analysis of metabolome data are available (Hirai et al., 2004; Kim et al., 2007). We previously applied BL-SOM to analyze the fold-change database for the amounts of metabolites obtained via time-course sampling of *Arabidopsis thaliana* cells after salt-stress treatment (Kim et al., 2007).

Here, we have evaluated the clustering and visualizing power of BL-SOM to assess the metabolite level data generated from an earlier time-course design. *Arabidopsis* T87 cells were obtained from the RIKEN Bio Resource Center (Tsukuba, Japan). They were grown in a modified liquid LS medium (30 mL) in a 100 mL flask at 23°C and under continuous light [photosynthetic photon flux density (PPFD) of 55 μmol m^{-2} s^{-1}]. After 3 d, the cells were treated with 100 mM NaCl, then sampled for analysis after 0.5, 1, 2, 4, 12, 24, 48, and 72 h. Control cells also were harvested by filtration, immediately ground in liquid nitrogen, and stored at –80°C. All metabolites were extracted, separated, identified, and measured via LC/MS/MS, LC/MS, and GC/MS analysis according to the procedures previously reported by our group (Kim et al., 2007).

Using the BL-SOM algorithm, we examined the data from 47 metabolites and calculated the ratios of *S*-adenosyl-L-methionine to *S*-adenosyl-L-homocysteine (methylation index) at eight time points during this salt-stress treatment (Table 1). Mean values for metabolite levels were taken from three replicates. This algorithm organized the data into a two-dimensional matrix by an iterative process that was based on the relative similarity of the fold-change patterns of these metabolites. Relative values for normalized metabolite levels

*Corresponding author; fax +82-31-299-1122
e-mail kjkpj@rda.go.kr

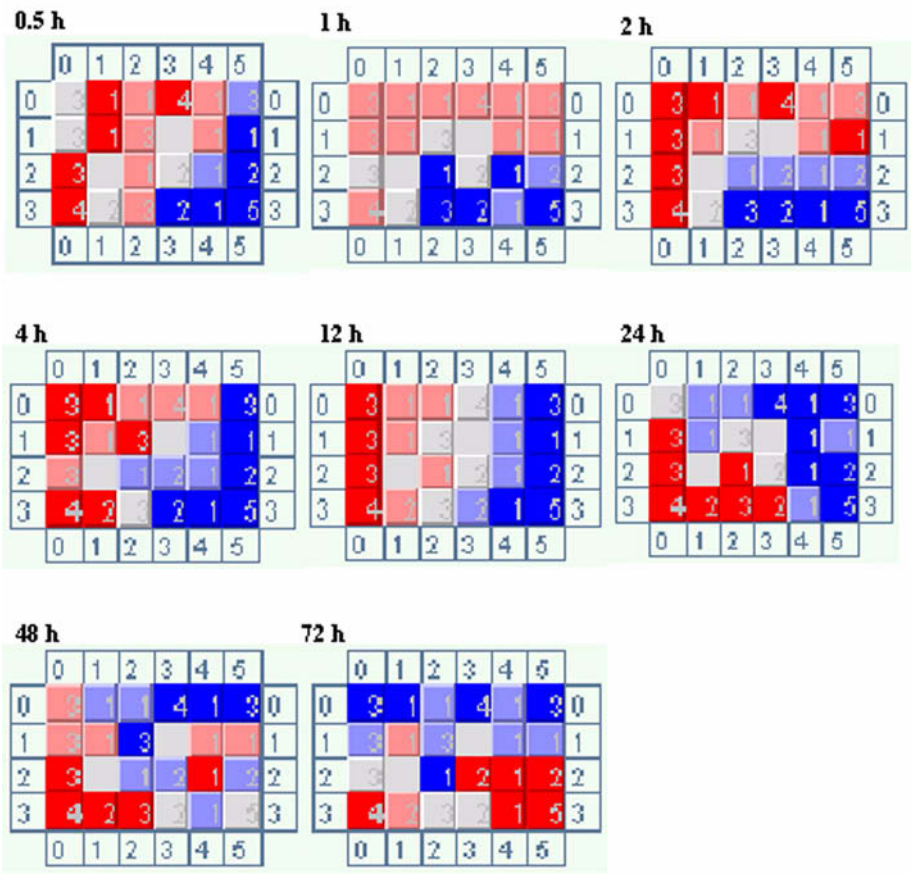

Figure 1. A 6 × 4 matrix for BL-SOM analysis on metabolomics. SOMs for metabolite data of differential response with respect to time series. Neurons were arranged in two-dimensional lattice. Metabolite numbers clustered are shown in neurons. Metabolite clusters are on 22 neurons, and comparative changes in metabolite levels are indicated by neuron colors: red (most increased), pink (increased), pale blue (decreased), and blue (most decreased).

were used, and the highest level attained by each over the time-course was set to '1'. At the beginning, each neuron of the SOM was assigned an initial weight vector defined by PCA. These weight vectors (w_{ij}) were arranged into a two-dimensional lattice denoted by i (=0, 1, …, I-1) and j (=0, 1, …, J-1), where I was set as '6', J was defined by the nearest integer greater than $(\sigma_2/\sigma_1) \times I$, and σ_1 and σ_2 were the standard deviations of the first and second principal components, respectively. During the learning phase, the Euclidian distance was calculated between one sample vector and all the weight vectors of the map. Thus, the weight vectors for the best matching neuron were moved toward the values of the input vectors, such that the neurons represented a group of similar accumulation profiles.

Here, BL-SOM rapidly and reliably clustered the metabolites into groups with similar concentration patterns, and which were metabolically related. Previously, we had performed an analysis in order to cluster metabolites with similar level profiles in the same neuron into a 4 × 3 matrix SOM, where the 24 metabolites were classified into 4 groups (Kim et al., 2007). In the present study, to produce better separation of the different patterns, all metabolites with similar level profiles were clustered within the same neuron and/or the neighboring neuron into a 6 × 4 matrix consisting of 22 neurons (Fig. 1). The number of metabolites in individual neurons varied between 1 and 5. Neurons with decreasing level profiles of metabolites appeared on the upper side and metabolites that were accumulated during the salt-stress treatment were located down and to the right. Metabolites that had no remarkable changes clustered to the lower-left side of the field. Except for Ser, Gly, Pro, Ala, and Glu, all the amino acids were contained in the upper-side neurons, while metabolites related to glycolysis metabolism clustered toward neurons on the lower-right side, suggesting that the clustering achieved by BL-SOM can reliably predict functional similarity (Fig. 2A, B).

Changes in metabolite levels were either dramatic or subtle. Those of the former kind were easily recognized, whereas the more subtle changes required statistical processing to determine whether they were significant. Line graphs for each neuron provided profound evidence of correlations between metabolite levels and the schemes that could reconstruct metabolic networks from the time series (Fig. 2A). For example, metabolite levels for the upper-right clusters -- Phe, Lys, Tyr, Trp, and Cys – were suppressed at the time point of 24 h before those of the neighboring cluster that contained ethanolamine, citric acid, gluconic acid, and the methylation index decreased. After that 24-h point, the metabolite levels decreased for the upper-left clusters, which included Val, Ile, Asn, Gln, and Leu, while those of

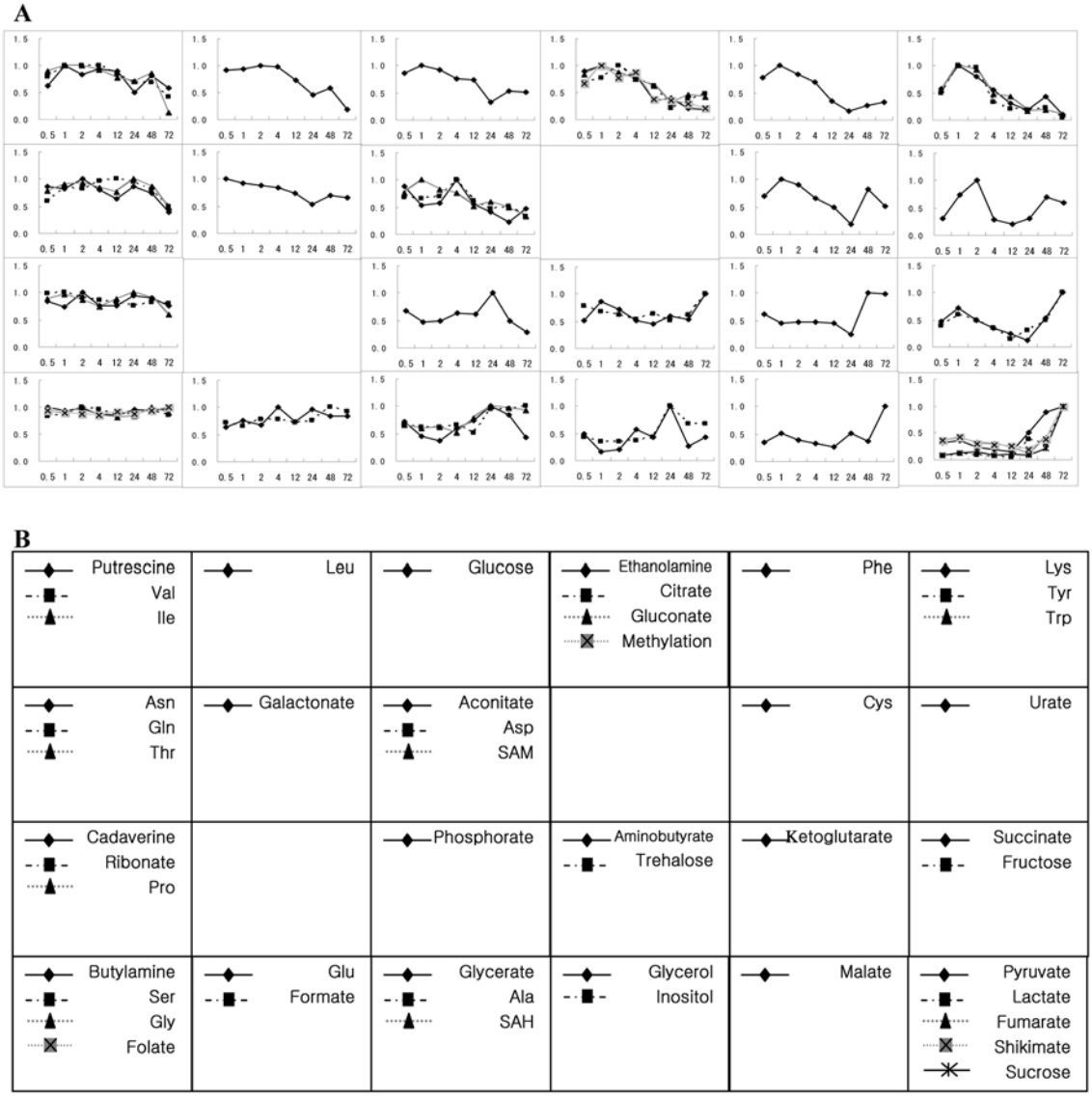

Figure 2. Line graphs (**A**) and metabolite clusters (**B**) displayed on 22 neurons (Fig. 1) obtained by BL-SOM analysis. Highest level for each metabolite in raw data (Table 1) was set to '1', which gave relative levels for each time-course sample. X axis represents time (h) of sampling and Y axis represents relative positioning of normalized metabolite levels on line graph (**A**). Comparative changes in metabolite levels are shown in boxes.

the lower-right clusters, e.g., α-ketoglutaric acid, succinic acid, D-fructose, malic acid, and pyruvic acid, were accumulated. Hence, it is most interesting that the metabolites related to glycolysis and sucrose metabolism were dramatically altered in conjunction with changes in amino acid content. Roessner et al. (2006) have reported that amino acid biosynthesis in potato tubers is regulated by sucrose levels. We also noted that levels of shikimic acid increased at 72 h, while those of Tyr, Trp, and Phe declined. Shikimate is produced by condensation of phosphoenolpyruvate and erythrose 4-phosphate, followed by several reactions. The decrease in Tyr and Phe seen here could also be interpreted as the degradation of those compounds. Evidence supporting this hypothesis includes the rise in fumaric acid, a breakdown product of aromatic amino acid catabolism. Moreover, among the amino acids analyzed, only Ala was increased at 72 h for T87 cell cultures in the presence of salt (Table 1). Under such stress conditions, Ala is accumulated as a storage form of pyruvate (Ben-Izhak Monselise et al., 2003).

In summary, based on our analysis using BL-SOM, we can suggest that a negative correlation exists between the metabolites associated with glycolysis metabolism and those amino acids that respond to salt stress. The clustering achieved by BL-SOM represents functional similarity, which reflects participation in a common or closely related pathway. Therefore, we conclude that this BL-SOM approach for cluster analysis could be applied for grouping similar patterns of metabolite levels, and could also be used to infer underlying biological mechanisms.

Table 1. Time-course change of metabolite levels in T87 cells treated with 100 mM NaCl. Data are normalized to the mean response calculated for the control cells of each measured batch. (SAM, *S*-adenosyl-L-methionine; SAH, *S*-adenosyl-L-homocysteine). Values are mean ± SEs (n=3).

Metabolite	30 min	1 h	2 h	4 h	12 h	24 h	48 h	72 h
Pyruvate	1.19 ± 0.08	1.32 ± 0.27	0.86 ± 0.00	0.68 ± 0.14	0.54 ± 0.15	1.84 ± 0.47	3.23 ± 0.63	3.63 ± 0.07
Lactate	0.87 ± 0.00	1.26 ± 0.31	0.98 ± 0.01	0.67 ± 0.26	0.53 ± 0.41	4.94 ± 1.45	3.67 ± 0.69	12.75 ± 0.25
Phosphorate	1.29 ± 0.01	0.92 ± 0.10	0.95 ± 0.01	1.21 ± 0.13	1.19 ± 0.09	1.92 ± 0.07	0.96 ± 0.13	0.55 ± 0.00
n-Butylamine	1.09 ± 0.15	1.08 ± 0.12	1.07 ± 0.02	0.96 ± 0.11	0.95 ± 0.19	1.05± 0.04	1.03 ± 0.05	1.07 ± 0.16
Ethanolamine	1.31 ± 0.04	1.45 ± 0.31	1.18 ± 0.02	1.22 ± 0.14	0.52 ± 0.01	0.59 ± 0.03	0.31 ± 0.02	0.26 ± 0.02
Glycerol	1.61 ± 0.06	0.72 ± 0.24	0.68 ± 0.02	1.91 ± 0.04	1.42 ± 0.09	3.30 ± 0.21	0.90 ± 0.07	1.40 ± 0.04
Succinate	0.75 ± 0.01	1.08 ± 0.12	0.80 ± 0.02	0.55 ± 0.04	0.39 ± 0.05	0.21 ± 0.00	0.87 ± 0.07	1.62 ± 0.01
Glycerate	1.17 ± 0.05	0.82 ± 0.12	0.58 ± 0.00	0.97 ± 0.07	1.18 ± 0.07	1.60 ± 0.17	1.35 ± 0.30	0.70 ± 0.04
Fumarate	0.82 ± 0.03	0.92 ± 0.10	0.72 ± 0.02	0.70 ± 0.02	0.53 ± 0.01	0.23 ± 0.06	0.91 ± 0.14	2.30 ± 0.05
Cadaverine	1.06 ± 0.04	1.00 ± 0.05	1.07 ± 0.04	0.89 ± 0.03	0.92 ± 0.08	0.91 ± 0.06	1.08 ± 0.04	0.96 ± 0.01
Malate	0.79 ± 0.03	1.11 ± 0.15	0.87 ± 0.06	0.73 ± 0.01	0.62 ± 0.03	1.17 ± 0.38	0.82 ± 0.06	2.27 ± 0.05
4-Aminobutyrate	0.87 ± 0.00	1.26 ± 0.24	1.21 ± 0.00	0.87 ± 0.05	0.77 ± 0.04	1.01 ± 0.07	0.92 ± 0.06	1.70 ± 0.11
Cysteine	1.03 ± 0.08	1.50 ± 0.18	1.34 ± 0.15	0.97 ± 0.08	0.74 ± 0.11	0.27 ± 0.05	1.23 ± 0.25	0.78 ± 0.60
α-Ketoglutarate	1.17 ± 0.17	0.99 ± 0.26	0.90 ± 0.17	0.91 ± 0.29	0.84 ± 0.27	0.47 ± 0.05	1.88 ± 0.42	1.84 ± 0.03
Aconitate	1.50 ± 0.11	0.91 ± 0.21	1.00 ± 0.03	1.71 ± 0.40	0.95 ± 0.18	0.70 ± 0.14	0.37 ± 0.04	0.82 ± 0.14
Putrescine	0.60 ± 0.01	0.94 ± 0.06	0.82 ± 0.02	0.92 ± 0.14	0.87 ± 0.06	0.49 ± 0.02	0.80 ± 0.13	0.56 ± 0.01
Ribonate	1.01 ± 0.01	1.00 ± 0.08	0.92 ± 0.02	0.87 ± 0.03	0.83 ± 0.04	0.78 ± 0.08	0.84 ± 0.01	0.81 ± 0.02
Shikimate	0.71 ± 0.08	0.90 ± 0.19	0.60 ± 0.02	0.54 ± 0.03	0.53 ± 0.01	0.37 ± 0.06	0.76 ± 0.04	2.02 ± 0.14
Citrate	1.10 ± 0.10	1.15 ± 0.25	1.60 ± 0.15	1.24 ± 0.07	1.00 ± 0.05	0.34 ± 0.10	0.61 ± 0.07	0.76 ± 0.02
D-fructose	0.72 ± 0.02	1.05 ± 0.10	0.89 ± 0.01	0.64 ± 0.05	0.26 ± 0.02	0.62 ± 0.04	0.85 ± 0.15	1.59 ± 0.09
Glucose	0.97 ± 0.00	1.02 ± 0.02	1.02 ± 0.01	0.85 ± 0.04	0.85 ± 0.03	0.38 ± 0.03	0.63 ± 0.01	0.58 ± 0.01
Lysine	1.32 ± 0.02	1.74 ± 0.62	1.85 ± 0.01	1.28 ± 0.23	0.70 ± 0.07	0.26 ± 0.08	0.97 ± 0.04	0.25 ± 0.01
Tyrosine	1.81 ± 0.02	3.71 ± 0.47	3.64 ± 0.21	1.25 ± 0.30	0.76 ± 0.06	0.26 ± 0.05	0.84 ± 0.34	0.14 ± 0.03
Gluconate	0.95 ± 0.01	1.08 ± 0.08	1.00 ± 0.02	0.85 ± 0.05	0.70 ± 0.04	0.36 ± 0.03	0.51 ± 0.04	0.48 ± 0.02
Galactonate	1.07 ± 0.07	0.94 ± 0.05	0.95 ± 0.00	0.90 ± 0.02	0.78 ± 0.03	0.56 ± 0.03	0.75 ± 0.05	0.71 ± 0.05
Inositol	1.03 ± 0.00	0.88 ± 0.13	0.83 ± 0.01	0.88 ± 0.05	1.00 ± 0.03	2.37 ± 0.24	1.63 ± 0.08	1.60 ± 0.03
Urate	0.76 ± 0.01	1.52 ± 0.45	2.50 ± 0.10	0.74 ± 0.19	0.52 ± 0.10	0.78 ± 0.25	1.76 ± 0.16	1.46 ± 0.04
Sucrose	0.86 ± 0.01	1.10 ± 0.11	1.49 ± 0.06	0.89 ± 0.15	1.08 ± 0.11	0.53 ± 0.04	2.12 ± 0.22	10.03 ± 0.05
Trehalose	1.11 ± 0.08	0.96 ± 0.13	0.88 ± 0.04	0.75 ± 0.09	0.90 ± 0.10	0.73 ± 0.06	0.87 ± 0.13	1.42 ± 0.02
Aspartate	0.77 ± 0.55	0.74 ± 0.09	0.80 ± 0.06	1.14 ± 0.17	0.70 ± 0.02	0.53 ± 0.06	0.59 ± 0.08	0.38 ± 0.05
Asparagine	1.01 ± 0.11	0.97 ± 0.02	1.17 ± 0.11	0.95 ± 0.06	0.74 ± 0.08	1.00 ± 0.02	0.87 ± 0.03	0.46 ± 0.07
Serine	0.79 ± 0.02	0.82 ± 0.04	0.94 ± 0.12	0.90 ± 0.01	0.82 ± 0.01	0.82 ± 0.02	0.88 ± 0.04	0.80 ± 0.00
Glycine	0.91 ± 0.01	0.85 ± 0.02	0.96 ± 0.01	0.82 ± 0.03	0.78 ± 0.01	0.81 ± 0.04	0.97 ± 0.03	0.87 ± 0.04
Alanine	0.92 ± 0.04	0.86 ± 0.09	0.85 ± 0.03	0.91 ± 0.04	0.71 ± 0.00	1.31 ± 0.06	1.30 ± 0.03	1.38 ± 0.06
Glutamate	0.46 ± 0.02	0.55 ± 0.02	0.48 ± 0.02	0.71 ± 0.03	0.52 ± 0.04	0.69 ± 0.07	0.59 ± 0.05	0.61 ± 0.05
Glutamine	0.45 ± 0.03	0.63 ± 0.08	0.61 ± 0.04	0.73 ± 0.03	0.75 ± 0.03	0.72 ± 0.02	0.60 ± 0.04	0.37 ± 0.04
Threonine	0.90 ± 0.06	1.04 ± 0.04	1.08 ± 0.06	0.98 ± 0.01	0.89 ± 0.06	1.16 ± 0.03	1.01 ± 0.08	0.54 ± 0.03
Proline	0.96 ± 0.03	1.04 ± 0.02	0.93 ± 0.02	0.80 ± 0.01	0.95 ± 0.03	1.08 ± 0.01	0.97 ± 0.03	0.64 ± 0.04
Valine	0.81 ± 0.01	1.01 ± 0.06	1.01 ± 0.11	1.01 ± 0.04	0.85 ± 0.03	0.69 ± 0.01	0.70 ± 0.01	0.43 ± 0.05
Trptophan	1.31 ± 0.06	2.34 ± 0.09	2.18 ± 0.05	1.11 ± 0.04	1.01 ± 0.07	0.37 ± 0.02	0.45 ± 0.03	0.24 ± 0.03
Isoleucine	1.07 ± 0.04	1.21 ± 0.03	1.20 ± 0.14	1.12 ± 0.03	0.92 ± 0.04	0.85 ± 0.05	1.04 ± 0.04	0.16 ± 0.02
Leucine	1.11 ± 0.01	1.15 ± 0.03	1.22 ± 0.17	1.20 ± 0.00	0.89 ± 0.04	0.56 ± 0.03	0.72 ± 0.02	0.22 ± 0.02
phenylalanine	1.47 ± 0.06	1.88 ± 0.08	1.60 ± 0.01	1.31 ± 0.05	0.66 ± 0.03	0.29 ± 0.00	0.50 ± 0.01	0.61 ± 0.10
Formate	1.02 ± 0.02	0.92 ± 0.03	1.10 ± 0.03	1.09 ± 0.01	1.01 ± 0.01	1.08 ± 0.02	1.41 ± 0.01	1.29 ± 0.01
SAM	1.38 ± 0.04	1.82 ± 0.02	1.47 ± 0.04	1.39 ± 0.07	0.92 ± 0.01	1.08 ± 0.08	0.89 ± 0.01	0.60 ± 0.05
SAH	1.09 ± 0.06	0.96 ± 0.06	1.02 ± 0.07	0.85 ± 0.06	1.29 ± 0.05	1.65 ± 0.04	1.60 ± 0.10	1.53 ± 0.05
Methylation	1.25 ± 0.05	1.88 ± 0.03	1.45 ± 0.05	1.63 ± 0.06	0.71 ± 0.02	0.65 ± 0.05	0.56 ± 0.05	0.39 ± 0.05
Folate	0.93 ± 0.06	0.90 ± 0.01	0.88 ± 0.01	0.87 ± 0.00	0.94 ± 0.02	0.89 ± 0.02	0.94 ± 0.01	1.01 ± 0.03

ACKNOWLEDGEMENT

The authors are grateful to Dr. Shigehiko Kanaya for providing the computer program for BL-SOM.

Received March 21, 2007; accepted July 11, 2007.

LITERATURE CITED

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. Genome Res 13: 693-702

Abe T, Sugawara H, Kanaya S, Kinouchi M, Ikemura T (2006) Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes. Gene 365: 27-34

Ben-Izhak Monselise E, Parola AH, Kost D (2003) Low-frequency electromagnetic fields induce a stress effect upon higher plants, as evident by the universal stress signal, alanine. Biochem Biophys Res Commun 302: 427-434

Fiehn O (2002) Metabolomics - the link between genotypes and phenotypes. Plant Mol Biol 48: 155-171

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol 18: 1157-1161

Fukusaki E, Kobayashi A (2005) Plant metabolomics: Potential for practical operation. J Biosci Bioengr 100: 347-354

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 101: 10205-10210

Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nishi T, Mori H, Ikemura T (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): Characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. Gene 276: 89-99

Kim JK, Bamba T, Harada K, Fukusaki E, Kobayashi A (2007) Time-course metabolic profiling in *Arabidopsis thaliana* cell cultures after salt stress treatment. J Exp Bot 58: 415-424

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43: 59-69

Kohonen T (1990) The self-organizing map. Proc IEEE 78: 1464-1480

Kohonen T, Oja E, Simula O, Visa A, Kangas J (1996) Engineering applications of the self-organizing map. Proc IEEE 84: 1358-1384

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13: 11-29

Roessner U, Urbanczyk-Wochniak E, Czechowski T, Kolbe A, Willmitzer L, Fernie AR (2003) De novo amino acid biosynthesis in potato tubers is regulated by sucrose levels. Plant Physiol 133: 683-692

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: Large-scale phytochemistry in the functional genomics era. Phytochemistry 62: 817-836

Tikunov Y, Lommen A, de Vos CH, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics: Large-scale profiling of tomato fruit volatiles. Plant Physiol 139: 1125-1137