Integrated Data Mining of Transcriptome and Metabolome Based on BL-SOM

Mitsuru Yano^{a,c}, Shigehiko Kanaya^b, Md. Altaf-Ul-Amin^b,
Ken Kurokawa^b, Masami Yokota Hirai^c, Kazuki Saito^{*a,c}

^a Graduate School of Pharmaceutical Science, Chiba University, Inage-ku, Chiba 263-8522, Japan,

^b Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0101, Japan. ^c RIKEN Plant Science Center, Yokohama 230-0045, Japan.

(Received July 10,2006; Accepted August 28, 2006)

With the recent advances of sophisticated bioinformatics tools and analytical technology, the systems biological approach becomes more realistic to solve biological problems. The aim of this study is to illustrate a feasibility of batch-learning self-organizing map (BL-SOM) for high-throughput analysis of post-genomic data on plant biology. BL-SOM is a modification of the conventional SOM, which provides colored feature maps independent of the order of data input. We conducted the nutritional stress experiments on a model plant *Arabidopsis thaliana* and applied the BL-SOM to analysis of the data sets of metabolome and transcriptome obtained by Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICRMS) and DNA microarray. BL-SOM analyses of non-targeted metabolome data provided metabolic fingerprinting of global responses to nutritional stress and developments. Using the clustering performance of BL-SOM, we deduced metabolite identification from accurate mass values and elucidated a metabolic mechanism responding to sulfur deficiency. Transcriptome data were also analyzed by BL-SOM, and genes responding to sulfur deficiency were classified according to their expression patterns. The result showed that functionally related genes were clustered in the same or neighbor lattice points. We examined each cluster and deduced putative functions of genes involved in glucosinolate biosynthesis, and the function of some of those genes was identified by biochemical experiments. Our present study suggests that application of BL-SOM to an integrated post-genomic omics data gives great possibilities for more accurate prediction in systems biology. The BL-SOM software is provided freely at our web site (<http://prime.psc.riken.jp>).

Key Words: batch-learning SOM, metabolome, transcriptome, data-mining, plant science

* ksaito@faculty.chiba-u.jp

1. Introduction

The transition from a reductionistic to a more holistic approach has occurred by progressing total genome sequencing and the complete transcriptional analysis of an organism. This is particularly the case concerning the newest field of functional genomics and metabolomics [1]. Metabolomics is the study of global metabolite profiles in a system such as cell, tissue, or organism under a given set of conditions, and has been labeled as one of the new “omics” joining genomics, transcriptomics, and proteomics, which are employed toward the understanding of global systems biology [2, 3].

The combination of improved analytical capabilities with bioinformatics is beginning to broaden the horizons of our understanding how metabolism is controlled and how it is highly flexible. In chemo- and bio-informatics, at least following four points are required for metabolomics; comparative analysis between different data sets; processing and mining the data to extract the components of interest; visualization strategies to easily understand present complex data; and databasing for efficient data storage [1].

Plants have typically been forced to evolve complex multicellular structures of divergent function and synthesize a wide variety of secondary metabolites in order to survive in a continually changing and predominantly hostile environment including abiotic (light, UV, nutrition, salinity and drought) and biotic (herbivore, parasitism and pathogen attack) stress factors. Non-targeted approach of metabolomics, which are in progress recently, is particularly suitable initial approach as a hypothesis generator to provide early leads for future research.

Self-organizing map (SOM) facilitates visualization and interpretation of large data sets [4-8], which is the reason why SOM is applied to understand multivariate data set in various fields. We recently developed a modification of the conventional SOM for genome and post-genome analyses, that makes the learning process and resulting map independent of the order of data input

[9,10]. In the present study, we apply batch-learning self-organizing map (BL-SOM) to integrated analysis of metabolome and transcriptome data sets, demonstrating that SOM makes comprehensive understanding possible by an example of holistic response to nutritional stress and time-series data in metabolite and transcript levels. The four case studies (CS1 to CS4) were carried out as following: [CS1] metabolic fingerprinting of global responses to nutritional stresses in *Arabidopsis*; [CS2] deduction of metabolite annotation from non-targeted metabolome data obtained by FT-ICRMS analysis; [CS3] deduction and confirmation of gene functions; and [CS4] experiment for integrated analysis of transcriptome and metabolome. Software utilized in the present study is downloadable from <http://prime.psc.riken.jp/>.

2. Experiments

The purpose of the present study is to demonstrate the way that one comprehensively understands high-throughput data sets produced by metabolome and transcriptome data based on BL-SOM. The procedure is outlined in Figure 1. BL-SOM was applied to metabolome data sets in CS1 and CS2, to transcriptome data set in CS3, to both metabolome and transcriptome data sets in CS4. In CS1, we used *Arabidopsis thaliana* which is a popular model plant in plant biology. We performed two series of nutritional stress experiments. In long-term stress response experiment, *Arabidopsis* plants were grown hydroponically in the media of control, sulfur deficient, nitrogen deficient, and both sulfur and nitrogen deficient for 17 days. In short-term stress response experiment, plants were grown on agarose-solidified control medium for 3 weeks, then transferred to the media of control, sulfur deficient, and *O*-acetyl-L-serine (OAS)-supplemented, and further grown for 48 hours. Leaves and roots were harvested separately. The details of sampling and analytical methods are described in the previous paper [11].

In CS2-4, *Arabidopsis* plants were grown under sulfur-deficient condition for 3, 6, 12, 24, 48 and 168

Figure 1 Flow chart of metabolome and transcriptome analysis by BL-SOM.

hours, and leaves and roots were harvested separately. Details of this time series experiments are described in the previous paper [12]. Metabolites were extracted from leaves and roots into polar- and non-polar-fractions, and each fraction was subjected to in-fusion mass spectrometric analysis. Electrospray ionization (ESI) (positive and negative) and atmospheric pressure chemical ionization (APCI) (positive and negative) modes were employed. Average signal intensity of triplicate analyses was calculated to each accurate m/z peak, which was calibrated carefully between all measurements. In CS3 and CS4, total RNA was extracted from leaves and roots separately, and transcriptomes were analyzed using DNA microarray as described [12]. Transcriptome data were normalized appropriately, and fold change of signal intensity in sulfur-starved sample to that in sulfur-sufficient control sample was calculated.

The algorithm of BL-SOM has previously been described [9, 10]. In this section, we explain how we use BL-SOM for data mining in metabolome and/or transcriptome data. Figure 2 shows the concept of Self-organizing map.

3.1. Learning process in BL-SOM

Initially, in s th object corresponding to metabolite/gene, metabolite-quantity/expression-level for t th experimental condition is denoted by x_{st} ($s=1, 2, \dots, N, t=1, 2, \dots, M$). So the metabolite/transcriptome data set consists of a matrix described by Eq.(1).

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1t} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2t} & \dots & x_{2M} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{s1} & x_{s2} & \dots & x_{st} & \dots & x_{sM} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nt} & \dots & x_{NM} \end{pmatrix} \quad (1)$$

3. How to use SOM for data mining

Figure 2 The concept of BL-SOM

The matrix data in Eq.(1) can be represented by the scatter distribution of objects (metabolites/genes) in M-dimensional space as shown in Figure 2(a). In learning process of SOM, the distribution of objects are reflected by the distribution of weight points $\mathbf{w}_{ij}$ ($i=1,2,\dots,I$; $j=1,2,\dots,J$) as shown in Figure 2(b). In the present study, we determined the initial weight vectors based on the first and second principal components of the M-dimensional space by principal component analysis (PCA). Weights in the first dimension (I) are arranged into lattices corresponding to a width that is five times the standard deviation ($5\sigma_1$) of the first principal component. The second dimension (J) is defined by the nearest integer greater than $(\sigma_2/\sigma_1) \times I$. The total number of weights in the first dimension I is set by a user. The weight vector on the ij th lattice ($\mathbf{w}_{ij}$) is calculated as follows:

$$\mathbf{w}_{ij} = \mathbf{x}_{av} + \frac{5\sigma_1}{I} \left[\mathbf{b}_1 \left(i - \frac{I}{2} \right) + \mathbf{b}_2 \left(j - \frac{J}{2} \right) \right]$$

Here $\mathbf{x}_{av}$ is the average vector for individual conditions of all input vectors, and $\mathbf{b}_1$ and $\mathbf{b}_2$ are eigenvectors for the first and second principal components.

The minimum Euclidean distance of any input vector $\mathbf{x}_k$ with respect to all weight vectors $\mathbf{w}_{ij}$ ($i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$) is determined and the corresponding weight vector is denoted by $\mathbf{w}_{i'j'}$. The input vector $\mathbf{x}_k$ is classified into the set S_{ij} for the lattice points (i, j) satisfying $i' - \beta(r) \leq i \leq i' + \beta(r)$ and $j' - \beta(r) \leq j \leq j' + \beta(r)$. After classification of all input vectors to the lattice points (i, j), weight vectors are updated by

Figure 3 Metabolic finger printings of various nutritional stress experiments. Number 1 to 8 are the results of long-term stress response in hydroponic culture, and 9 to 14 are those of short-term stress response in agarose-solidified plate culture. Number 1 to 4 and 9 to 11 are the results of root samples, and number 5 to 8 and 12 to 14 are those of leaf samples. cont, control condition; dS, sulfur deficiency; dN, nitrogen deficiency; dSdN, sulfur and nitrogen deficiency; OAS, OAS-supplemented.

$$\mathbf{w}_{ij}^{(new)} = \mathbf{w}_{ij} + \alpha(r) \left(\frac{\sum_{\mathbf{x}_k \in S_{ij}} \mathbf{x}_k}{N_{ij}} - \mathbf{w}_{ij} \right).$$

The two parameters $\alpha(r)$ and $\beta(r)$ are learning coefficients for the r th cycle, and N_{ij} is the number of components of S_{ij} . $\alpha(r)$ and $\beta(r)$ are calculated as follows:

$$\alpha(r) = \max \{0.01, \alpha(1)(1 - r/T)\}$$

$$\beta(r) = \max \{0, \beta(1) - r\}$$

Here, $\alpha(1)$ and $\beta(1)$ are the initial values for the T-cycle of the learning process. The learning process is monitored by the total distance between $\mathbf{x}_k$ and the nearest weight vector $\mathbf{w}_{ij}$, represented as

$$Q(r) = \sum_{k=1}^N \left\{ \left\| \mathbf{x}_k - \mathbf{w}_{i'j'} \right\|^2 \right\}$$

Here N is the total number of input vectors analyzed.

3.2. Feature Maps

In BL-SOM, we can easily visualize transcription or metabolite profiles for individual experimental conditions. Each of the objects is classified into lattice point whose distance is the minimum from the vector of object as shown in second layer of Figure 2(c). So the objects classified into an identical lattice point correspond to very similar transcript or metabolite profiles for all experimental conditions. In the context of individual experiments, lattice points were classified into three types (corresponding to third to last layers in Fig. 2(c)). In the case that all the intensities of objects included in the (i,j) -lattice point are larger than the average intensity for t th experiment ($t=1,2,\dots,M$), this lattice point is classified in high level (assigned by red),

Figure 4 BL-SOM analysis on the metabolome of leaf samples responding to sulfur deficiency. Glucosinolates, isothiocyanates and galactolipids formed clusters. The number in each lattice point represents number of peaks obtained by FT-ICRMS analysis.

Figure 5 Chemical structures of glucosinolates and isothiocyanates. Glucosinolates are hydrolyzed by thioglucosidase upon plant damage. Depending on the structure, the hydrolyzed products yield different compounds including isothiocyanates, nitriles, epithionitriles and thiocyanates. Abbreviations: Glu, glucose; R, various side chains.

otherwise, in the case that all the intensities in the (i,j) -lattice point are smaller than average intensity, this lattice point is classified in low level (assigned by blue). So, individual maps corresponding to experiments are colored by the intensity level of the profiles of objects, which are referred to as feature maps. When we compare feature maps between experimental conditions (Fig.2(c)), we can visually detect difference of expression or metabolite profiles in individual experiments. This is very effective to detect state transition points for time series data because transition points for expression or metabolite profiles are visually detected by dissimilarity of colored maps.

4. Case studies

4.1 Metabolic fingerprinting of global responses to nutritional stresses (CS1)

Plants take up sulfur as sulfate (SO_4^{2-}) and assimilate it into various organic compounds. Sulfur-containing metabolites play fundamental roles in maintenance of protein structure and in basic biochemical processes such as the energy-transfer reaction and the redox control [13]. Sulfur deficiency in environment is a critical stress for plants. In face of nutritional deficiency, plants must change gene expression pattern, metabolism process and growth rate to adapt themselves to the stressed conditions. Plant response to sulfur deficiency has

Figure 6 Accumulation profiles of glucosinolates, isothiocyanates and galactolipids. Metabolites were putatively identified based on elemental composition calculated using mean mass value (the measured m/z value) and database searching. Abbreviations: MGDG, monogalactosyl diacylglycerol; DGDG, digalactosyl diacylglycerol; SQDG, sulfoquinovosyl diacylglycerol. Numbers following indicate acyl species; for example 18:3 means octadecatrienoate. Profiles are depicted only for the metabolites which correspond to those in the natural product database (Dictionary of Natural Products, John Buckingham, Chapman & Hall/CRC, England, CRC Press) by precise molecular weights.

been studied from various points of view [11, 12, 14-19]. To elucidate a holistic response of sulfur deficiency in plants, we have conducted the nutritional stress experiments with short-term and long-term time scales. Non-targeted metabolome data were obtained by FT-ICRMS, and metabolic fingerprinting of each experiment was produced by BL-SOM (Figure 3) with these data. The data set consists of metabolite quantities detected by approximately 3000 peaks by in-fusion FT-ICRMS analysis for 14 experimental conditions. Figure 3 shows feature maps on the metabolome of nutritional stress experiments. The results indicated that plant organ (root and leaf), method of culture (hydroponic and agarose-solidified plate) and period of stress (17days and 48 hours) were dominant factors to metabolite fingerprinting. Nutritional conditions (sulfur deficiency, nitrogen deficiency, and both sulfur and nitrogen deficiency) only slightly affected metabolite fingerprinting.

4.2 Deduction of metabolites from non-targeted metabolome data obtained by FT-ICRMS analysis (CS2)

Time series data of metabolite accumulation in *Arabidopsis* responding to sulfur deficiency were obtained by FT-ICRMS without pre-separation by chromatography. Fold change of ion intensity of each

accurate m/z peak was calculated as the ratio of average signal intensity in sulfur-starved samples to that in control samples. Each peak had six fold change values at 6 time points, and it was considered to be a vector in six-dimensional space. BL-SOM analysis was conducted with the data set consisting of these vectors. Using the clustering performance of BL-SOM, we could easily distinguish isotopic molecular ion peaks and sodium-/potassium-adducted molecular ion peaks from principal molecular ion peaks, because the vectors of isotopic molecular ion peaks and sodium-/potassium-adducted molecular ion peaks were almost same as that of principal molecular ion peaks, and so they were clustered with the principal molecular ion peaks. We eliminated them from the data set for the further processing. For each peak elemental composition was assigned based on accurate m/z within the error range of 5 ppm. Metabolites were annotated based on the elemental composition and subsequent database searching of natural products. For metabolite annotation, we took into account the presumption that metabolites showing similar patterns in change can be the intermediates/derivatives of the same metabolic pathway.

Approximately 850 peaks were detected in leaf samples under sulfur deficiency. BL-SOM analysis was conducted with this data set (Figure 4). We found that the accumulation of glucosinolate-related metabolites was coordinately changed under this condition.

Figure 7 (a) BL-SOM analysis on the transcriptome of leaf sample responding to sulfur deficiency. The known genes involved in glucosinolate biosynthesis (shown in blue letters) were classified into two clusters shown in pale green. Three putative sulfotransferase genes (shown in red letters) were classified into one of them. The other candidate genes involved in glucosinolate biosynthesis are shown in orange letters. MAM, methylthioalkylmalate synthase; CYP, cytochrome P450; S-GT, S-glucosyltransferase; GST, glutathione-S-transferase. (b) Representative pathway of glucosinolate biosynthesis. The formation of glucosinolate consists of two phases. First, side chains of certain amino acids are elongated, and second, amino acid moiety is transformed into the core structure of glucosinolate. Sulfotransferases participate in the last step of the second phase.

Glucosinolates are sulfur-rich natural products unique to Brassicaceae plants. They produce isothiocyanates upon hydrolysis by endogenous thioglucosidases released upon an attack by herbivore, and the degradation products act as defense compounds [20-21] (Figure 5). In the BL-SOM feature map, seven molecular species of glucosinolates were classified in the lattice points of (0, 0) and (0, 1), while isothiocyanates were classified in the lattice points of (12, 3) and (14, 2). Some of other peaks in these clusters were those of potassium or sodium ion adducts of glucosinolates or isothiocyanates (Fig. 6).

Accumulation profiles of glucosinolates and isothiocyanates showed temporal change in reverse manner of each other (Fig. 6), agreeing with the conversion of isothiocyanates formation concomitant with glucosinolates decomposition. This result indicated that, at least during early stage of response to sulfur deficiency (3-6 hr), glucosinolates were decomposed to supply sulfate to essential pathways of sulfur metabolism, and biosynthesis of glucosinolates might be repressed because of acute shortage of sulfur. In middle stage (12-48 hr), glucosinolates accumulated again, presumably because defense response might become activated, and sulfate became available from other sulfur-containing metabolites. In the later stage (168 hr), levels of glucosinolate decreased by a long-term sulfur deprivation.

Many peaks detected by FT-ICRMS were assigned to lipids in our non-targeted metabolome analysis of sulfur-starved samples. Among them, the peaks assigned to galactolipids formed a cluster (Fig. 4). Galactolipids are of most abundant lipid class in leaves, which contain galactolipids abundantly in chloroplast lamellae [22]. Accumulation profiles of galactolipids changed temporally responding to sulfur deficiency (Fig. 6). It suggested that lipid metabolism was involved in sulfur-deficient response. This speculation could be supported by the fact that sulfate is mainly reduced in plastids [23]. Sulfur deficiency affects redox and photosynthesis in chloroplasts, resulting in changes of lipid composition of membrane in chloroplasts. In addition, Nikiforova *et al.* (2003) [17] reported that UDP-sulfoquinovose:DAG sulfoquinovosyltransferase gene (*SQD2*) involved in biosynthesis of sulfolipids, which are major components of galactolipids and exclusively associated with plastid membranes [24], was down-regulated in their sulfur depletion study using *Arabidopsis*.

4.3 Elucidation of gene functions (CS3)

Non-targeted transcriptome data was obtained with the same plant samples used in the metabolome analysis. We conducted BL-SOM analysis with the transcriptome data set consisting of signal log ratio to each gene at six time points, and deduced the functions of genes which

could be involved in glucosinolate biosynthesis [12].

Among 21,500 genes on a DNA microarray, approximately 10,000 genes were significantly expressed and used in the subsequent analysis. Figure 7 (a) shows the result of BL-SOM on the transcriptome data set. Many known genes involved in glucosinolate biosynthesis were classified only into two clusters. In one of these clusters, there were three putative genes which were presumed to code for sulfotransferases according to their sequence homology. We speculated that these genes could code for sulfotransferase committed in the last step of glucosinolate biosynthesis, because the genes were not identified till then. *In vitro* enzymatic assay was performed using the recombinant proteins produced by these genes, confirming their function as sulfotransferase in glucosinolate biosynthesis [12].

4.4 Experiment for integration analysis of transcriptome and metabolome (CS4)

When an appropriate normalization is done, transcriptome and metabolome data obtained from the same plant material can be integrated into a single multivariate data set suitable for further BL-SOM analysis. Genes and metabolites which change synchronously are classified into the same cluster. In addition various ways of integration are possible. For example, before integration of two data sets, the plus/minus sign of one data set can be reversed, then we can obtain negatively correlated items to each other. Shifting of the time point of one data set to early or late time point before integration is also possible, and in this way we can find time-delay correlation. These modifications of analysis would give us more clues for gene and metabolite identification.

Figure 8 shows an example of such modification with the integrated data set of root samples. In this case we normalized the different types of data sets as follows; the metabolites and transcripts, whose maximum absolute values of log-base10 ratio were smaller than 0.1 through time course, were eliminated. For ca.1,000 metabolites and ca.10,000 transcripts left after the elimination, the sum of the square of the six log ratio values at six time points was set equal to 1 to give relative log ratio values, and all data were combined into a single data set. This normalization procedure enabled us to extract closely-related metabolites and genes in sulfur-starved *Arabidopsis*. In *Arabidopsis* 36 genes are considered to be involved in sulfur assimilation pathway from sulfate to cysteine (Figure 9). Figure 8 (b) shows the result of BL-SOM analysis using the data set in which the sign of transcriptome data was reversed and integrated with the metabolome data. In the resulting map, sulfur assimilation genes and sulfate were clustered together. It indicates that, among genes involved in sulfur

assimilation, expression of sulfate transporter genes (*Sultr1:1*, *Sultr1:2*, *Sultr2:1*, *Sultr4:1*), 5'-adenylylsulfate reductase genes (*APR2*, *APR3*), ATP-sulfurylase gene

(*APS3*), and serine O-acetyltransferase genes (*Serat3:1*, *Serat3:2*) [25] was induced coordinately with decrease of sulfate. It is well known that sulfate assimilation is

Figure 8 BL-SOM analysis on the integrated data of transcriptome and metabolome. (a) The result of simply integrated data set. (b) The result of integrated data set in which the sign of gene expression data was reversed. In this map the metabolites and the genes which correlate negatively to each other are classified into the same cluster.

Figure 9 Sulfate assimilation pathway in *Arabidopsis*. Abbreviations: OAS, *O*-acetylserine; APS, adenosine 5'-phosphosulfate; APR, APS reductase; PAPS, 3'-phosphoadenosine 5'-phosphosulfate.

up-regulated under sulfur starvation and by high demand for sulfur metabolites [13]. It was reasonable that intracellular concentration of sulfate showed negative correlation to the amount of transcripts of genes involved in sulfate assimilation. On the other hand, when BL-SOM analysis was conducted using the data set which was obtained by simply combining the transcriptome and the metabolome data, these genes and *O*-acetylserine, precursor of cysteine, were classified together (Figure 8 (a)). Hirai *et al.* [19] reported that exogenous application of *O*-acetylserine mimics sulfur-deficiency stress.

5. Conclusions and Remarks

The advantageous point of BL-SOM compared to other multivariate analyses is its predictive power to biological problems via classification of unknown elements such as genes and metabolites. Our present study indicates that BL-SOM is a powerful methodology for gene identification from the transcriptome and metabolome data, especially for genes involved in biosynthesis of secondary metabolites. Gene-to-metabolite correlation is generally clear in secondary metabolism compared to primary metabolism [12, 26].

Metabolomics has emerged and become a predominant sector of post-genomics with the trend of systems biology. Metabolomics is predicted to play a significant role in understanding the relationship between genome and phenotype in cells and organisms [1-3]. Non-targeted metabolome analysis by FT-ICRMS is a developing technique and it provides suggestive information on holistic state of organisms. FT-ICRMS analyzer has the highest resolving power among mass spectrometers currently available, and it allows high-throughput analyses of extremely complex mixtures without chromatographic separation [27, 28]. However, for identification of metabolites by FT-ICRMS, biological information such as co-occurrence of analogous metabolites in the same pathway is additionally required. BL-SOM is very suitable to analyze such metabolome data obtained from FT-ICRMS, because the BL-SOM neighborhood lattice points will be able to predict intermediates, derivatives, ion-adducts and other chemically or biologically relating compounds. Though only a part of all peaks is successfully identified at the present stage of metabolomics, complete identification of metabolites would be possible accompanying with the improvements of database, identification procedure and methods for sample preparation in future.

Application of BL-SOM to metabolomics enables not only high-throughput analysis of metabolome data but also that of integrated omics data obtained from various phases in biology. If we have higher quality omics data

obtained at closer time points, BL-SOM analysis with integrated omics data would provide more accurate predictions.

Acknowledgments

The part of this research was supported in part by Core Research for Evolutional Science and Technology (CREST) of Japan Science and Technology Agency, by Grant-in-Aids for Scientific Research from the Ministry of Education, Science, Culture, Sports, and Technology, Japan, and by Uehara Memorial Foundation, Japan.

References

- [1] R. D. Hall, *New Phytologist*, **169**, 453-468 (2006).
- [2] K.-M. Oksaman-Caldentey and K. Saito, *Curr. Opin. Biotech.*, **16**, 174-179 (2005).
- [3] K. Saito, R. A. Dixon and L. Willmitzer *eds*: *Biotechnology in Agriculture and Forestry 57 Plant Metabolomics*, Springer Verlag, Heidelberg, (2006).
- [4] S. Mavroudi, S. Papadimitriou and A. Bezerianos, *Bioinformatics*, **18**, 1446-1453 (2002).
- [5] L. Xiao, K. Wang, Y. Teng and J. Zhang, *FEBS Lett.*, **538**, 117-124 (2003).
- [6] P. Mangiameli, S. K. hen and D. West, *Eur. J. Operational Res.*, **93**, 402-417 (1996).
- [7] S. Kaski, J. Nikkila, M. Oja, J. Venna, P. Toronen and E. Castren, *BMC Bioinformatics*, **4**:48, 1-13 (2003).
- [8] M. Vracko, *Curr. Comput-Aided Drug Design*, **1**, 73-78 (2005).
- [9] S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori and T. Ikemura, *Gene*, **276**, 89-99 (2001).
- [10] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki and T. Ikemura, *Genome Res.*, **13**, 693-702 (2003).
- [11] M. Y. Hirai, M. Yano, D. B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito, *Proc Natl Acad Sci USA* **101**, 10205-10210 (2004).
- [12] M. Y. Hirai, M. Klein, Y. Fujikawa, M. Yano, D. B. Goodenowe, Y. Yamazaki, S. Kanaya, Y. Nakamura, M. Kitayama, H. Suzuki, N. Sakurai, D. Shibata, J. Tokuhisa, M. Reichelt, J. Gershenzon, J. Papenbrock, and K. Saito, *J. Biol. Chem.*, **280**, 25590-25595 (2005).
- [13] K. Saito, *Plant Physiol.*, **136**, 2443-2450 (2004).
- [14] M. J. Hawkesford, *J Exp Bot.*, **342**, 131-138 (2000).
- [15] L. Hopkins, S. Parmar, A. Blaszczyk, H. Hesse, R. Hoefgen, and M. J. Hawkesford, *Plant Physiol.*, **138**, 433-440 (2005).
- [16] A. Maruyama-Nakashita, E. Inoue, A. Watanabe-Takahashi, T. Yamaya, and H. Takahashi, *Plant Physiol.*, **132**, 597-605 (2003).

- [17] V. Nikiforova, J. Freitag¹, S. Kempa, M. Adamik¹, H. Hesse and R. Hoefgen, *Plant J.*, **33**, 633-650 (2003).
- [18] V. J. Nikiforova, J. Kopka, V. Tolstikov, O. Fiehn, L. Hopkins, M. J. Hawkesford, H. Hesse, and R. Hoefgen, *Plant Physiol.*, **138**, 304-318 (2005).
- [19] M. Y. Hirai, T. Fujiwara, M. Awazuhara, T. Kimura, M. Noji and K. Saito, *Plant J.*, **33**, 651-663 (2003).
- [20] U. Wittstock and B. A. Halkier, *TRENDS in Plant Science*, **7**, 263-270 (2002).
- [21] B. A. Halkier and J. Gershenzon, *Annu. Rev. Plant Biol.*, **57**, 303-333 (2006).
- [22] J. Joyard, E. Teyssier, C. Miege, D. Berny-Seigneurin, E. Marechal, M. A. Block, A-J. Dorne, N. Rolland, G. Ajlani and R. Douce, *Plant Physiol.*, **118**, 715-723 (1998).
- [23] T. Leustek and K. Saito, *Plant Physiol.*, **120**, 637-643 (1999).
- [24] C. Benning, *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **49**, 53-75 (1998).
- [25] C. G. Kawashima, O. Berkowitz, R. Hell, M. Noji and K. Saito, *Plant Physiol.*, **137**, 220-230 (2005).
- [26] T. Tohge, Y. Nishiyama, M. Y. Hirai, M. Yano, J. Nakajima, M. Awazuhara, E. Inoue, H. Takahashi, D. B. Doodenowe, M. Kitayama, M. Noji, M. Yamazaki, and K. Saito, *Plant J.*, **42**, 218-235 (2005).
- [27] S. C. Brown, G. Kruppa, and J-L. Dasseux, *Mass Spectrometry Reviews*, **24**, 223-231 (2005).
- [28] Aharoni A, de Vos CHR, Verhoeven HA, Maliepaard CA, Kruppa G, Bino RJ and Goodenowe DB., *OMICS* **6**, 217-234 (2002).

一括学習型自己組織化マップ(BL-SOM)を利用したメタボロームおよびトランスクリプトームデータの統合解析

矢野美弦^{a,c}、金谷重彦^b、Md. Altaf-Ul-Amin^b、黒川顕^b、平井優美^c、
齊藤和季^{*a,c}

^a千葉大学大学院薬学研究院 遺伝子資源応用研究室、^b奈良先端科学技術大学院大学 情報科学研究科、^c理化学研究所 植物科学研究センター

本論文の目的は、我々が開発した一括学習型自己組織化マップ (BL-SOM) のソフトウェアを活用することにより植物科学の分野で得られるポストゲノム科学の多変量データから生物学的に有用な知見を効率よく抽出できることを例証することである。BL-SOMは典型的なSOMを特に分子生物学の分野におけるデータ解析に向けて改変したものであり、学習の結果がデータの入力順に依存しないという特性を持っているため結果の再現性が保証されている。我々はこの解析ツールをフーリエ変換イオンサイクロトロン型質量分析装置 (FT-ICRMS) による代謝物蓄積量の網羅的測定 (メタボローム) とマイクロアレイによる遺伝子発現解析 (トランスクリプトーム) を行うポストゲノムの手法による植物の栄養ストレス応答に関する研究に応用した。メタボロームデータの解析では、まずBL-SOMの可視化機能により植物の栄養ストレス応答の全体像をメタボリックフィンガープリンティングとして直感的に把握できる様式で示すことができ、植物が栽培条件や組織に依存して代謝物蓄積量プロファイルを大きく変化させていることを明らかにした。また、BL-SOMのもうひとつの特性である優れたクラスタリング機能を利用して、精密質量数から生物学的にも合理的と考えられる化合物名を推定することができ、欠乏させた元素を含む重要な二次代謝物の蓄積量が経時的に大きく変化することを明らかにした。トランスクリプトームデータの解析においては、適切な時系列データをBL-SOMで解析することにより、その機能が塩基配列からの推定にとどまっている遺伝子の機能を複数の代謝経路において同時に効率よく包括的に予測する方法を確立した。さらにこの方法によって予測した遺伝子の機能のいくつかを酵素学的方法によって確認することができた。BL-SOMはポストゲノムにおけるomics科学の分野で大量に得られる異なった種類のデータを統合して解析する際にも極めて有用な知識抽出のためのツールとなる可能性を示した。本ソフトウェアは以下のサイトにおいて無償でダウンロードできる (<http://prime.psc.riken.jp>)。

キーワード: 一括学習型自己組織化マップ (BL-SOM)、メタボローム、トランスクリプトーム、データマイニング、植物科学

* ksaito@faculty.chiba-u.jp