BMC Systems Biology

BioMed Central

Research article

Open Access

Time-resolved metabolomics reveals metabolic modulation in rice foliage

Shigeru Sato[1], Masanori Arita[1,2,3], Tomoyoshi Soga[1,4], Takaaki Nishioka[1,5] and Masaru Tomita*[1,4]

Address: [1]Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan, [2]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo and PRESTO-JST, Kashiwa, Japan, [3]Plant Science Center, Riken, Yokohama, Japan, [4]Human Metabolome Technologies, Inc., Tsuruoka, Japan and [5]Graduate School of Agriculture, Kyoto University, Kyoto, Japan

Email: Shigeru Sato - n03615ss@sfc.keio.ac.jp; Masanori Arita - arita@k.u-tokyo.ac.jp; Tomoyoshi Soga - soga@sfc.keio.ac.jp; Takaaki Nishioka - takaaki@sfc.keio.ac.jp; Masaru Tomita* - mt@sfc.keio.ac.jp

* Corresponding author

Published: 18 June 2008

BMC Systems Biology 2008, **2**:51 doi:10.1186/1752-0509-2-51

Received: 12 February 2008
Accepted: 18 June 2008

This article is available from: http://www.biomedcentral.com/1752-0509/2/51

Abstract

Background: To elucidate the interaction of dynamics among modules that constitute biological systems, comprehensive datasets obtained from "omics" technologies have been used. In recent plant metabolomics approaches, the reconstruction of metabolic correlation networks has been attempted using statistical techniques. However, the results were unsatisfactory and effective data-mining techniques that apply appropriate comprehensive datasets are needed.

Results: Using capillary electrophoresis mass spectrometry (CE-MS) and capillary electrophoresis diode-array detection (CE-DAD), we analyzed the dynamic changes in the level of 56 basic metabolites in plant foliage (*Oryza sativa* L. ssp. *japonica*) at hourly intervals over a 24-hr period. Unsupervised clustering of comprehensive metabolic profiles using Kohonen's self-organizing map (SOM) allowed classification of the biochemical pathways activated by the light and dark cycle. The carbon and nitrogen (C/N) metabolism in both periods was also visualized as a phenotypic linkage map that connects network modules on the basis of traditional metabolic pathways rather than pairwise correlations among metabolites. The regulatory networks of C/N assimilation/dissimilation at each time point were consistent with previous works on plant metabolism. In response to environmental stress, glutathione and spermidine fluctuated synchronously with their regulatory targets. Adenine nucleosides and nicotinamide coenzymes were regulated by phosphorylation and dephosphorylation. We also demonstrated that SOM analysis was applicable to the estimation of unidentifiable metabolites in metabolome analysis. Hierarchical clustering of a correlation coefficient matrix could help identify the bottleneck enzymes that regulate metabolic networks.

Conclusion: Our results showed that our SOM analysis with appropriate metabolic time-courses effectively revealed the synchronous dynamics among metabolic modules and elucidated the underlying biochemical functions. The application of discrimination of unidentified metabolites and the identification of bottleneck enzymatic steps even to non-targeted comprehensive analysis promise to facilitate an understanding of large-scale interactions among components in biological systems.

Background

In the post-genome era, comprehensive data from "omics" technologies (genomics, transcriptomics, proteomics, and metabolomics) have been extensively analyzed to elucidate the underlying biochemical networks that elaborately regulate cellular mechanisms. Recent contributions from metabolomics are particularly noteworthy; they offer insights into metabolism that complement information obtained from proteomics and transcriptomics [1]. Correlation analysis of metabolic profiles has been used effectively to distinguish silent phenotypes or genetic alterations that are not noticeable superficially [2-4]. The systematic integration of metabolomic-, proteomic-, and transcriptomic profiles facilitates the unbiased, information-based reconstruction of underlying biochemical networks [5,6]. Kohonen's self-organizing map (SOM) analysis [7] was also an effective method to classify and monitor metabolic alteration patterns with time-series profiles [8,9].

However, with the current technology, unbiased reconstruction from comprehensive and high-throughput data is challenging; statistical tools are immature and inherent measurement errors and biological noise continue to present problems [10]. Moreover, two issues are relevant to the exploitation of metabolomics data. First, it is crucial to interpret metabolic profiles by focusing on a specific rhythm in an appropriate time range and interval, since plants have adapted their metabolism to different environmental fluctuations such as the slow and steady diurnal rhythm, whereas metabolic levels change dynamically. Second, currently available metabolomics data are insufficient for the detection of new metabolic networks. Even if non-target profiling were able to quantify thousands of metabolites, at present there is no method for estimating their reliability. As statistical inference requires large amounts of data measured under similar conditions in transcriptomics [11], the verification of network dynamics for known pathways must precede attempts to identify unknown network structures. It appears that each metabolic profile is measured under method-specific, presumably biased conditions.

Time-resolved target analysis is an effective way to observe biochemical dynamics. We systematically measured the level of 56 basic metabolites in rice leaves (*Oryza sativa* L. ssp. *japonica*) at hourly intervals over a 24-hr period. Our target and experimental conditions were strategically determined: 1) we focused on primary metabolic pathways consisting of carbon fixation/respiration- and nitrogen assimilation/dissimilation pathways, and comprehensively quantified related metabolites, 2) the photocycle was the sole environmental factor, and 3) measurements were made at 1-hr intervals to allow the observation of dynamic profiles.

High-throughput analysis was conducted with the capillary electrophoresis – mass spectrometry (CE-MS) technology we developed earlier [12-14], and has been applied to metabolic profiling in *Bacillus subtilis* extracts [15] and monitoring of genetic and environmental perturbations in *Escherichia coli* cells [16]. Each employed CE-MS method was able to detect charged low molecular metabolites in less than 30 min without requiring derivatization. Combined with diode array detection (CE-DAD), our technology is also applicable to quantifying small sugar compounds. We previously developed a sample preparation protocol that could extract metabolites with possibly minimal metabolic turnover [17]. By using the CE-MS and CE-DAD, we also succeeded in analyzing over eighty major metabolites (sugars, organic acids, amino acids, and nucleotides) in rice foliage. The current work is our first systematic time-course measurements of rice foliage throughout a day.

We applied four information-based methods to analyze the diurnal fluctuation of metabolites: 1) metabolic pathways were classified with SOM to monitor the metabolic dynamics in each time-step, 2) a phenotypic linkage map was constructed from the classified pathways by Sammon's 2D-network layout [18], 3) unidentified metabolites were predicted based on SOM analysis and chemical structures, and 4) rate-limiting enzymes were identified by hierarchical clustering on a correlation matrix. Here we show that combining metabolome analysis and information-based methods is an effective way to elucidate phenotypical metabolic network structures and underlying biological functions under diurnal rhythm fluctuations.

Results

Time-course data acquisition

We extracted target metabolites existing in the primary metabolism such as the glycolytic pathway, the reductive- and oxidative pentose phosphate pathway, and the photorespiratory pathway, the tricarboxylic acid (TCA) cycle, and the amino acid biosynthetic pathway. Figure 1 presents the practical rice biochemical network that was constructed with our target metabolites based on annotated protein data from the KEGG pathway database [19], Swiss-Prot database [20], or Rice Annotation Project Data Base [21]. It shows the names of target metabolites and the EC number of enzymatic reactions; black dots are non-target metabolites. Although NH_3 (also R-NH_2) and CO_2 were non-target compounds, they are shown in green to demonstrate in and out of carbon and nitrogen.

We selected eight enzymatic proteins that have not been annotated at this stage to determine whether they function in the rice plant. These enzymes and the judgment criteria are shown in Table 1. On the map, their EC numbers and lines are presented in gray.

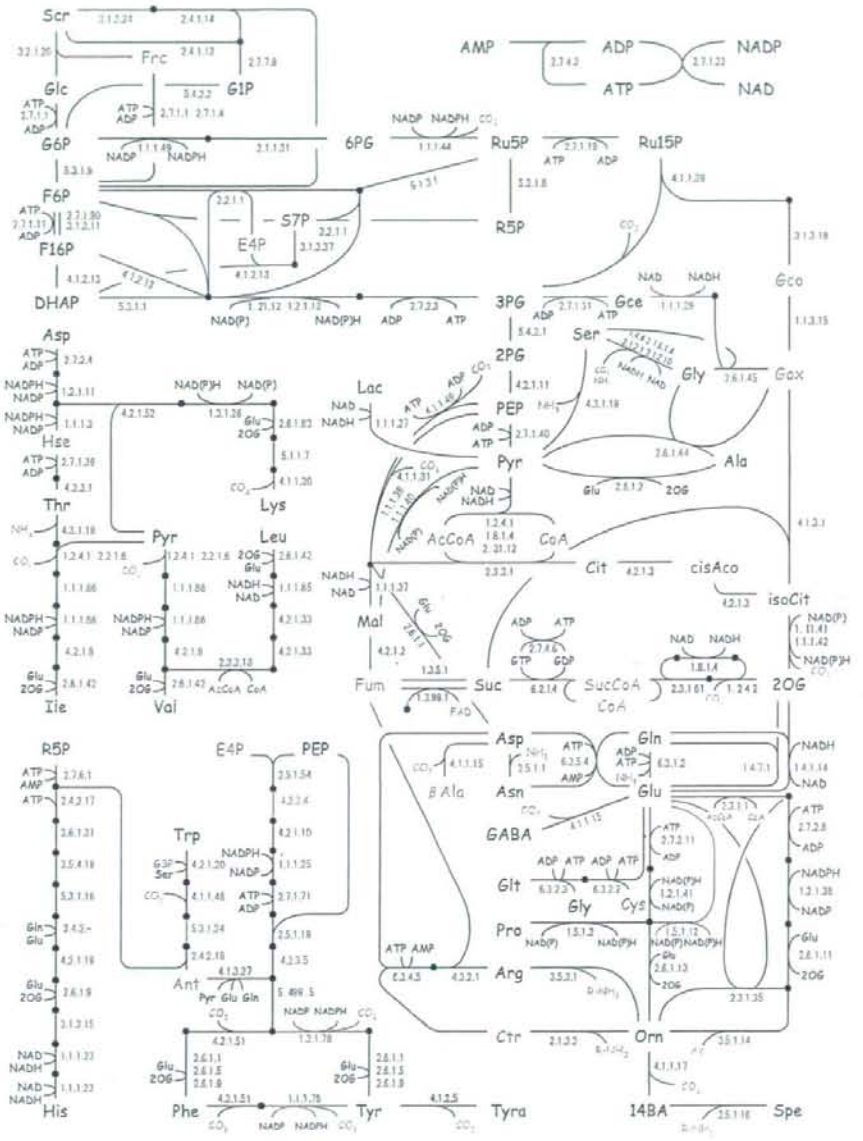

Figure 1

Metabolic network of *oryza sativa* L. ssp. *Japonica*. Target metabolites and practical enzymatic reactions are shown. The number next to the line is the EC number. Colors indicate the ratio of metabolic levels in light and dark periods. Unidentified metabolites are gray and gray lines and EC numbers identify non-annotated enzymatic proteins. The red- and yellow shade show the glycolytic pathway and the TCA cycle respectively.

Table 1: Selected non-annotated proteins expected to function in rice plant

EC Number	Enzyme name	Criterion for judgement	Ref.
1.1.1.29	hydroxypyruvate reductase; glycerate dehydrogenase	Enzymatic reduction of hydroxypyruvic acid to D-glyceric acid in higher plants, i.e. the leaves of pea, beet, tomato, radish, spinach, parsley, lettuce, corn, kohlrabi, and carrot.	[22]
		AK069655; Similar to 2-hydroxyacid dehydrogenase	RAP-DB*1
1.2.1.13	glyceraldehyde-3-phosphate dehydrogenase	AK071685; Similar to GADPH (383AA) (Fragment). AK67755; Similar to Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.13) (Fragment).	RAP-DB
1.3.1.78	arogenate dehydrogenase; prehenate dehydrogenase	TyrAAT1(AF434681) and TyrAAT2(AF434682) in *Arabidopsis thaliana* catalyze the oxidative decarboxylation of arogenate into Tyr in the presence of NADP. TyrAAT also exhibits prephenate dehydrogenase activity.	[23]
		Q5Z9H5_ORYSJ; Q5Z9H3_ORYSJ; Q5Z6Y1_ORYSJ, Putative arogenate dehydrogenase isoform 2	Swiss-Prot/TrEMBL*2
1.5.1.12	delta-1-pyrroline-5-carboxylate dehydrogenase	AK121765; Similar to delta-1-pyrroline-5-carbozylate dehydrogenase	RAP-DB
2.7.1.31	D-glycerate 3-kinase	GLYK family protein was purified and sequenced from *Arabidopsis thaliana*, identified as putative kinase-annotated single-copy gene At1g8038. This article suggests that an *Oryza sativa* PRK/UK-like protein, BAD73764, Os01g48990 is grouped with the GLYK kinase family.	[24]
3.1.3.24	sucrose-phosphatase	AK063330, AK071525, AK064563; Similar to sucrose-phosphatase	RAP-DB
4.2.3.4	3-dehydroquinate synthase	Pentafunctional aroma enzyme in Saccharomyces cervisiae includes EC 4.2.3.4, EC 4.2.1.10, EC 2.5.1.19, EC 1.1.1.25, and EC 2.7.1.71.	[25]
		AK071977; Similar to 3-dehydroquinate synthase-like protein (EC 4.2.3.4). Four other proteins were annotated.	RAP-DB
5.3.1.24	phosphoribosyl-anthranilate isomerase	J075072K08; Similar to phosphoribosylanthranilate isomerase	RAP-DB

*1: Rice Annotation Project Data Base [21]
*2: UniProt Knowledge base: Swiss-Plot and TrEMBL [20]

Sedoheptulose 1,7-bisphosphate (S17P) in the pentose phosphate pathway was not identified because the standard reagent was unavailable. Xylulose 5-phosphate (X5P) is a stereoisomer of Ribulose 5-phosphate (Ru5P) and their peak overlap in CE-MS analysis makes the identification even more difficult. Glyceraldehyde 3-phosphate (G3P) and oxaloacetate (OAA) were not accurately determined too, because they were readily reacted or decomposed.

The seventy selected target metabolites were classified into four groups according to their chemical structure-based physiochemical characteristics (Table 2). Group A contained amino acids and amines, group B organic acids and sugar phosphates, group C nucleotides and coenzymes, and group D sugars. Groups A, B, and C, consisting of ionic substances, were analyzed with three CE-MS methods for cationic, anionic, and nucleotide metabolites; analysis of group D was with a CE-DAD method. For CE separation, we used conventional sample preparation with simple and universal procedures without any derivatization process. As common preparation procedures were applicable under the four analytical conditions, we were able to determine simultaneously a wide variety of chemical compounds.

Plant seedlings were grown under a 13-hr light – 11-hr dark photocycle for 20 to 21 days. The level of the 56 metabolites was successfully quantified at hourly intervals over the course of 24 hr. We could identify the peak and determine the peak area for S7P but could not quantify its level, since the reagent was not available at the time of our CE-MS measurement; we later qualitatively identified its peak with the migration time ratio (MT/MT_{IS}) of S7P to PIPES (internal standard). The other 13 metabolites were under the detection limit (signal-to-noise ratio (S/N) < 3); their names were colored gray in Figure 1.

In the course of 24 hr, the metabolites exhibited various fluctuations (Figure 2). Ru15P, the precursor of carbon fixation, manifested a variation synchronous with the photoperiod; its intracellular concentration increased under illumination and decreased in darkness. Several metabolites exhibited similar light-dependent variations in the reductive pentose phosphate pathway (3PG, R5P, and Ru5P), the glycolytic pathway (3PG, 2PG, PEP, Pyr), the TCA cycle (2OG, Suc, and Mal), and in sugars (Scr and Glc). Citrate, on the other hand, manifested opposite fluctuation changes. In the amino acid biosynthesis pathway, major amino acids (Ala, Asn, Gln, Glu, Gly, and Ser) accumulated during the light period. Minor amino acids that

Table 2: The 70 target metabolites subjected to analysis of time-resolved dynamics and their abbreviation used in this article

Group A (CE-MS No.1)		Group B (CE-MS No.2)		Group C (CE-MS No.3)	
Amino acids		***Organic acids***		***Nucleotides***	
Ala	Alanine	cisAco	cis-Aconitate	AMP	AMP
β Ala	β-Alanine	Cit	Citrate	ADP	ADP
GABA	γ-Aminobutyrate	isoCit	iso-Citrate	ATP	ATP
Ant	Anthranilate	DHAP	Dihydroxyacetonephosphate	GDP	GDP
Arg	Arginine	Fum	Fumarate	GTP	GTP
Asn	Asparagine	Gce	Glycerate	***Coenzymes***	
Asp	Aspartate	Gco	Glycolate	NAD	NAD
Ctr	Citrulline	Gox	Glyoxylate	NADH	NADH
Cys	Cysteine	Lac	Lactate	NADP	NADP
Glu	Glutamate	Mal	Malate	NADPH	NADPH
Gln	Glutamine	2OG	2-Oxoglutarate	CoA	CoA
Glt	Glutathione red.	PEP	Phosphoenolpyruvate	AcCoA	Acetyl-CoA
Gly	Glycine	6PG	6-Phosphogluconate	SucCoA	Succinyl-CoA
His	Histidine	2PG	2-Phosphoglycerate		
Hse	Homoserine	3PG	3-Phosphoglycerate		
Leu	Leucine	Pyr	Pyruvate	Group D (CE-DAD)	
Ile	iso-Leucine	Suc	Succinate	***Sugars***	
Lys	Lysine	***Sugar Phosphate***		Frc	Fructose
Orn	Ornithine	E4P	Erythrose 4-phosphate	Glu	Glucose
Phe	Phenylalanine	F16P	Fructose 1,6-bisphosphate	Suc	Sucrose
Pro	Proline	F6P	Fructose 6-phosphate		
Ser	Serine	G1P	Glucose 1-phosphate		
Thr	Threonine	G6P	Glucose 6-phosphate		
Trp	Tryptophan	R5P	Ribose 5-phosphate		
Tyr	Tyrosine	Ru15P	Ribulose 1,5-bisphosphate		
Val	Valine	Ru5P	Ribulose 5-phosphate		
Amines		S7P	Sedoheptulose 7-ohosphate		
14BA	1,4-Butanediamine				
Spe	Spermidine				
Tyra	Tyramine				

are synthesized from specific organic acids through several reaction steps (His, Ile, Leu, Lys, Phe, Trp and Val) accumulated during the dark period.

Table 3 shows the status of adenine nucleosides and nicotinamide coenzymes in the light and dark periods. Whereas the ratios of ADP, NADP, and NADH were almost equal in the light and dark periods, the ratios of AMP and NADPH were higher and those of ATP and NAD were lower in the light period (see Discussion).

Self-organizing map and phenotypic linkage of metabolic modules

To visualize the functioning networks throughout a 24-hr period, we classified the metabolites according to similarities in their time-dependent behavior by using Kohonen's self-organizing map (SOM) and Sammon's 2D-network layout (Sammon map). The time-dependent levels of each metabolite were represented as a 24-dimensional vector. On the SOM, the 57 metabolites were classified into a 24 × 24 lattice on the basis of vector similarity. The map was roughly divided into two major groups (see the dark gray line in Figure 3A). Metabolites with high levels in the light period are in the left area; those with high levels in the dark period are on the right in the map. On the SOM, each group was further classified and assigned to subgroups consisting of nitrogen- and carbon-assimilating compounds. Certain amino acids were arranged near their precursor organic acids, e.g., Glu/2OG. Gly, Ser, and Ala were grouped with synthetic pathway intermediates such as Pyr and Gce. The degree of similarity among metabolites was quantitatively visualized on the Sammon map; it shows approximate distances between metabolites on the SOM according to the Euclidean distance of the input vectors (Figure 3B). When we merged neighboring metabolites on the Sammon map we obtained 12 subsets of metabolites. Each subset is composed of metabolites that exhibit synchronous, time-dependent fluctuations, a "metabolic module". Metabolites in the same module were often neighbors in a traditional metabolic pathway network. Products that accumulated during the light period were arranged in subsets M1 – M8. They included the module for the reductive pentose phosphate pathway (M3), the photorespiratory pathway (M2), the latter half of the gly-

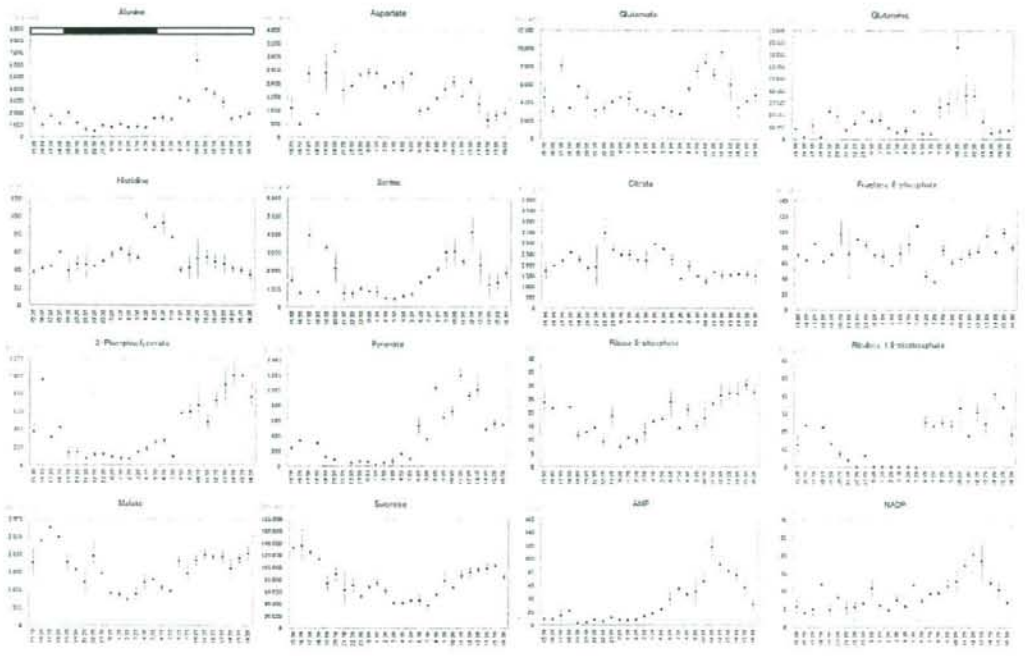

Figure 2
Metabolic time-courses in rice foliage at the third-leaf stage. Plantlets were grown under a 13-hr light – 11-hr dark photocycle. We applied 3 CE-MS methods and a CE-DAD method to analyze 69 major metabolites. Dynamic changes in the metabolite levels were assessed at hourly intervals over a 24 h period. Averages of 2 samples (± SEM) are shown. The top bar (shown in only Ala) indicates light and dark conditions.

colytic pathway (M4), the latter half of the TCA cycle (M5), sugars (M7), and major amino acids (M1). Also included in this group were NADPH and NADH (M6), glutathione and spermidine (M8). Subsets M9 – M12 included the first half of the glycolytic pathway (M9), the first half of the TCA cycle (M10), and minor amino acids (M11); also included were the nucleoside tri- and diphosphates (M12). Thus, our SOM analysis correctly reflected the phenotypic metabolic variations that indicate functioning biochemical pathways, and therefore represents a phenotypic linkage map (PLM).

Table 3: Status of adenine nucleosides and nicotinamide coenzymes in the light and dark period

	ATP AdN[*1]	ADP AdN	AMP AdN	NAD NiC[*2]	NADH NiC	NADP NiC	NADPH NiC
Light[*3]	0.21	0.40	0.40	0.36	0.10	0.09	0.44
Dark[*4]	0.45	0.43	0.11	0.55	0.09	0.05	0.31

[*1] AdN = ATP + ADP + AMP
[*2] NiC = NAD + NADH + NADP + NADPH
[*3] The average of all data throughout the light period
[*4] The average of all data throughout the dark period

The advantages of this analysis became even more apparent upon time-resolved analysis of metabolite levels (Figure 3C), which allowed visualization of the dynamic activity of these metabolic modules (see Discussion).

Discussion

Estimation of unidentified metabolites with SOM analysis

Although S17P could not be directly identified, we hypothesized that its peak could be identified in CE-MS data by combining SOM analysis with knowledge of the chemical structure. We identified a candidate peak among several peaks on selected ion electropherograms using a simple estimation method. As electrophoretic mobility is proportional to the ionic charge of the solute and inversely proportional to the size of the ionic molecule related to the hydrated ionic radius of a spherical mole-

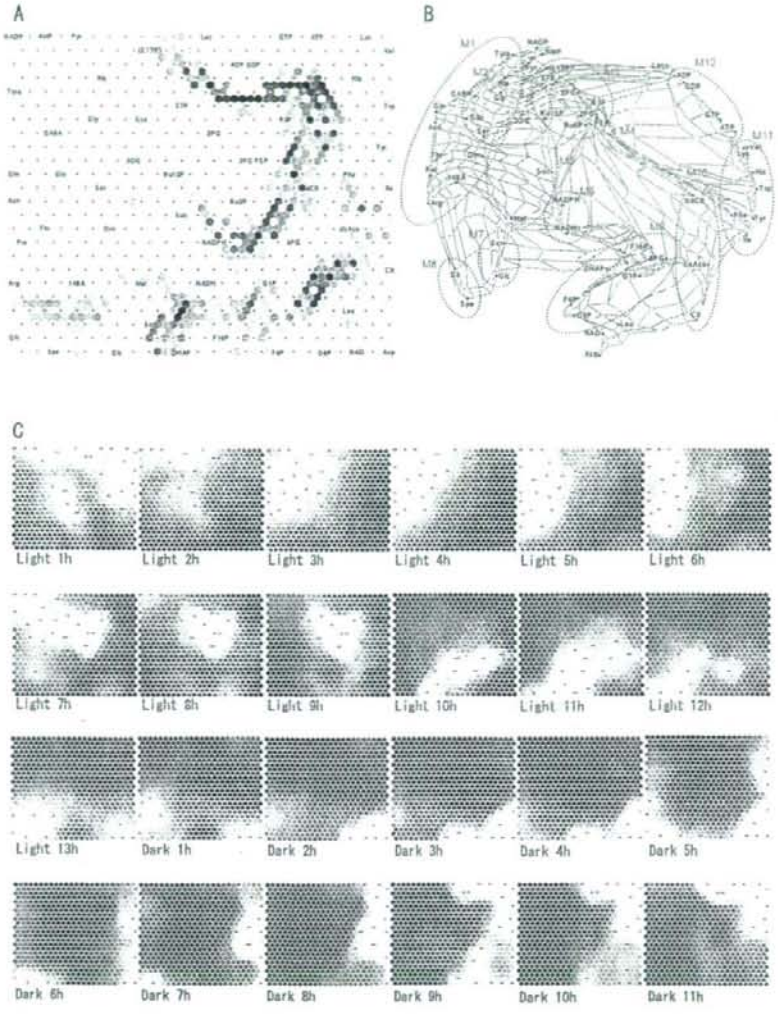

Figure 3

Self-organizing map (SOM) Analysis. A. U-matrix. Measured metabolites (n = 56) were arranged in a 20 × 20 lattice on the basis of diurnal change similarities. Light- and dark shading indicate high and low similarity, respectively. **B.** Phenotypic linkage map (PLM). The linkage among metabolites based on dynamic similarity is expressed as the distance on the quadratic plane. The metabolites were assigned to 14 metabolic modules that fluctuated synchronously; most contained traditional metabolic pathway networks or similar compounds. M1, major amino acid; M2, related to photorespiratory pathway intermediates; M3, pentose phosphate pathway; M4, latter half of the glycolytic pathway; M5, latter half of the TCA cycle; M6, environmental stress response; M7, sugars; M8, NADH and NADPH; M9, first half of the glycolytic pathway; M10; first half of the TCA cycle; M11, minor amino acids; M12, nucleoside tri- and diphosphates. **C.** Time-resolved layout. The relative levels of metabolites are shown for every time point from the start of the light period to the end of the dark period. Light and dark shading indicate high and low levels.

cule [26], we used the cubic root of the molecular weight as a substitute parameter for the radius. Indeed, the cubic root of molecular weights of 3 metabolites of similar chemical structure, Ru5P, F6P and S7P, were linearly correlated with migration time ratios ($r > 0.999$), when PIPES was used as an internal standard (Table 4).

The estimate for S17P was performed using linear approximation with Ru15P and F16P. The estimated migration time ratio (MT/MT_{IS}) of S17P was 0.941 (Table 4). Several peaks were observed at a mass-to-charge ratio (m/z) of 369. A peak of $MT/MT_{IS} = 0.909$ (m/z = 369) was identified within ± 5.0% of the predicted values.

Next, the absence of other metabolites with similar chemical structures was verified with the KEGG ligand database [27]. Note that except for S17P, metabolites were cyclic or non-anionic compounds.

Finally, we obtained the normalized time-course of the putative S17P by calculating the ratio of the peak area of putative S17P to PIPES. Integration of these data into the SOM analysis showed that this putative S17P marker was near metabolites in the reductive pentose phosphate pathway (Figure 3A) or the metabolic module M3 in PLM.

Unfortunately, the above result includes some speculation; most peaks of putative S17P were below the detection limits ($S/N < 3$) and the peak was not detected in the dark period. In the SOM analysis, the peak area of such undetected metabolite was calculated as zero. Nevertheless, the proposed estimation method seems to be effective in identifying unknown metabolites.

Detection of metabolic bottlenecks by pair-wise correlation analysis

In previous studies, Peason's correlation coefficients of metabolite pairs (pair-wise correlation) were applied to construct a metabolic correlation network [5,10,28]. A correlation coefficient is an index of co-linearity between two variables. If two metabolites, A and B, are always equilibrated, i.e., $[A]/[B] = K_{eq}$ (constant), then their relationship is linear and shows a high correlation. Although real metabolic pathways are dynamic and constantly regulated by their influx and/or efflux, the pathway components that are blocked by rate-limiting enzymes should exhibit approximate linearity. For example, 3PG, 2PG, and PEP in the glycolytic pathway are positioned between two rate-limiting enzymes, phosphoglycerate kinase (EC 2.7.2.3) and pyruvate kinase (PK; EC 2.7.1.40), both of which are regulated by the ATP/ADP ratio (Figure 1). The correlation coefficients among these three metabolites throughout a 24-hr period were over 0.90, whereas the correlation coefficient between PEP and Pyr, limited by PK, was under 0.50. Thus, pair-wise correlation analysis is effective for the identification of metabolic modules that are regulated by rate-limiting enzymes.

We used a hierarchical clustering algorithm, Ward's method [29], to classify metabolites in the glycolytic pathway (Figure 1) on the basis of their correlation matrix that was computed using all data throughout the 24-hr period. Indeed, a dendrogram identified the steps regulated by the ATP/ADP ratio (Figure 4A). On the other hand, it did not identify phosphofructokinase I (PFK-1; EC 2.7.1.11) as a rate-limiting enzyme. Although it is regulated by the ATP/ADP ratio in animal cells, another enzyme, pyrophosphate fructose 6-phosphate 1-phosphotransferase (EC 2.7.1.90), seems to be active in plant cells and may be independent of the ATP/ADP ratio [30].

The same cluster analysis was also applied to the TCA cycle intermediates (Figure 1), and the dendrogram revealed the rate-limiting enzymes in the cycle again (Figure 4B): citrate synthase (CS; EC 2.3.3.1), and NADP-dependent isocitrate dehydrogenase (ICDH; EC 1.1.1.42). This suggests that the classification of metabolites along enzymatic steps can help to reveal bottleneck enzymes.

Time-resolved carbon/nitrogen metabolomics

Inspection of the time-course of metabolic modules allowed us to better understand the carbon and nitrogen (C/N) assimilation/dissimilation process and their underlying function during a 24-hr period (Figure 3C).

In the first half of the light period, some accumulation emerged for carbon-fixed products: Pyr, 2OG, and photorespiratory pathway intermediates (metabolic module

Table 4: Estimated migration-time of unidentifiable metabolites based on the molecular weight of similar metabolites

Compound	Formula	M.W.	$M.W.^{1/3}$	MT/MT_{IS}
Ru5P	$CH_2(OH)CO\,[CH(OH)]_2CH_2OPO_3H_2$	230.0192	6.127	1.029
F6P	$CH_2(OH)CO\,[CH(OH)]_3CH_2OPO_3H_2$	260.0298	6.383	1.080
S7P	$CH_2(OH)CO\,[CH(OH)]_4CH_2OPO_3H_2$	290.0403	6.619	1.125
Ru15P	$CH_2(OPO_3H_2)CO\,[CH(OH)]_2CH_2OPO_3H_2$	309.9854	6.768	0.847
F16P	$CH_2(OPO_3H_2)CO\,[CH(OH)]_3CH_2OPO_3H_2$	339.9960	6.980	0.895
S17P	$CH_2(OPO_3H_2)CO\,[CH(OH)]_4CH_2OPO_3H_2$	370.0065	7.179	0.941*

*Estimated value. MT/MT_{IS} was calculated by linear approximation

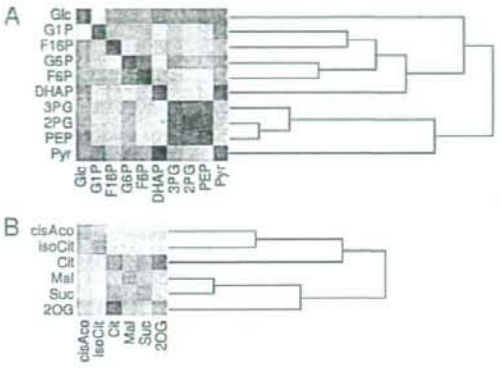

Figure 4
Hierarchical cluster analysis. A. Cluster analysis (Ward's method [26]) was applied to the correlation matrix composed of metabolic intermediates in the glycolytic pathway. The generated dendrogram was clustered into regulatory units by the ATP/ADP ratio; hexokinase (EC 2.7.1.1), phosphoglycerate kinase (EC 2.7.2.3), and pyruvate kinase (EC 2.7.1.40). **B.** As well as in the TCA cycle, the dendrogram was divided into two major groups at the rate-limiting steps; citrate synthase (CS; EC 2.3.3.1), and NADP-dependent isocitrate dehydrogenase (ICDH; EC 1.1.1.42).

M2). This coincides with carbon fixation by activation of several light-dependent enzymes including rubisco (EC 4.1.1.39) at the start of light exposure [31], as shown by the accumulation of Ru15P, Gce and triose derivatives at the beginning of the light period (light 1 – 3 hr). The slow accumulation was partly attributable to the very slow metabolic turnover of rubisco [32]. Likewise, major amino acids and amines including Glu and Gln, the source compounds of nitrogen assimilation as amino-group acceptor/donor [33,34], also accumulated in the first half of the light period (M1). This coincides with the diurnal metabolic dynamics and the activities of key enzymes in tobacco plant [35]. For example, NR activity is known to remarkably increase immediately after the start of light exposure and decrease at midday.

On the other hand, the glycolytic pathway and the reductive pentose phosphate pathway intermediates reached their highest levels (M3, M4) at midday, and sugars peaked at the end of the light period (M7).

We can hypothesize that carbon fixed in the first half of the light period moves down the glycolytic pathway and the TCA cycle, and amino acid biosynthesis progresses using generated Glu, Pyr, and 2OG. In the latter half of the light period, the flow of fixed carbon leads to the accumulation of the intermediates in the pentose phosphate pathway and to sucrose synthesis by inhibiting the production of ammonia, Pyr, and 2OG.

From the end of the light period through the first half of the dark period, we noted an increase in sugar phosphates from the first half of the glycolytic pathway (metabolic module M9). Around midnight, the accumulation of a few organic acids in the first half of the TCA cycle (metabolic module M10) was observed, suggesting the activation of the TCA cycle.

In the latter half of the dark period, the level of minor amino acids was increased (metabolic module M11), although they are synthesized from diverse biochemical pathways. The good correlation among these minor amino acids, also reported in potato and wheat [36], is attributable to the fact that the ratio between Gln and 2OG regulate minor amino acids in bacteria and fungi through the reaction Glu + 2-oxo acid ↔ amino acid + 2OG [37]. Under our experimental conditions, the Glu/2OG ratio was much higher in the dark- than in the light period (22.9 vs. 7.2) and the amino group can easily transferred to 2-oxo acids to produce amino acids.

Adenine nucleoside and nicotinamide coenzyme status

ATP and ADP were placed in the dark-activated group in PLM (metabolic module M12); they were accumulated at the end of the dark period, and decreased by illumination (Figure 3C). On the other hand, AMP was placed in the light-activated group peaking at midday. The reason for fluctuations of adenylate is unknown. Previous observations also do not coincide in the adenylate levels during the light- and dark period. In sugar beet leaves, all adenylate levels increased in the light period [38]. In spinach leaves and wheat leaf protoplast, ATP increased but ADP and AMP decreased under light [39,40]. In Crassulacean-acid metabolism (CAM) species, on the contrary, ATP decreased but ADP and AMP increased [41]. Such differences may result from different dynamics in cytosol, chloroplasts, and mitochondria [40].

We extrapolate that the lower ATP ratio during the light period was caused by an excess demand of ATP by intra- and extra cellular processes for carbon fixation and nitrogen assimilation against ATP supply from photosynthesis. In theory, the amount of ATP consumption in the reductive pentose phosphate pathway and the photorespiratory pathway is more than ATP production in the photophosphorylation [42]. Beside this, nitrogen assimilation process, intracellular transport of the assimilation products, and sucrose synthesis and its translocation are also accompanied by ATP. Therefore the dark respiration makes a considerable contribution to produce ATP even in the light. However, granted that ATP supply is insufficient in the light, high metabolic turnover of adenylate

kinase (EC 2.7.4.3) would immediately work to reproduce ATP from ADP that leads to increase of AMP. Further investigation is necessary to clarify the adenylate dynamics among cell compartments.

In our analysis, NADPH and NADH behaved similarly (metabolic module M6), whereas NADP and NAD did not. As NADPH and NADH were respectively generated by their unique reaction of reducing NADP and NAD, dependence on the intracellular oxidation-reduction state shifted the formation of oxidation and reduction. In PLM, however, NADP was placed in the light-activated- and NAD in the dark-activated group. This suggests that highly concentrated NAD in the dark is converted to NADPH via NADP in the light period. It was reported that the NADPH/NAD ratio is the inverse of the ATP/ADP ratio in guard cell protoplast, which indicates that ATP phosphorylates NAD in the light period by NAD kinase (EC 2.7.1.23) and the generated NADP is reduced to NADPH in the course of photosynthesis [43].

The ratios of NADH to NAD and NADPH to NADP were 0.16–0.29 and 6.2–6.6. The observed difference in the tendency of oxidized- or reduced form indicates their different cellular roles. NADH is used for oxidative phosphorylation, and a low NADH/NAD ratio constrains this process. On the other hand, NADPH is used for the reductive biosynthesis of metabolites, and the high ratio of NADPH/NADP favors the reduction of metabolites.

Environmental stress response

It is remarkable that Glt (GSH; gamma-glutamylcysteinyl glycine) and Spe exhibited similar fluctuation patterns (metabolic module M8). Both peaked at the end of the light period and again just after midnight, suggesting the existence of common regulatory factors. GSH plays a central role in the antioxidant defense by eliminating harmful peroxide during photosynthesis and oxidative phosphorylation [44]. Polyamines, including spermidine, are also effective antioxidants under various environmental stress conditions [45]. During photosynthesis, GSH is converted to oxidized dithiol (GSSH) to eliminate oxidative stress, and upon the reduction of NADPH, GSSH can be converted back to GSH by glutathione reductase (GR; EC 1.8.1.7, annotated in rice plant). Our finding that NADPH reached its highest level at a few hours before the end of the light period is consistent with the above observation (Figure 3C), although the connection remains speculative. The relative contribution of NADPH and NADH to the generation of GSH and spermidine requires further investigation.

Conclusion

We intended to analyze the rice plant metabolism and to reconstruct its phenotypic networks in an effort to explain underlying biological functions. Our CE-MS technology provided a comprehensive high-throughput system with easy sample preparation and facilitated the generation of high-resolution metabolic time-courses. Data mining with statistical techniques and SOM analysis revealed synchronous dynamics in metabolic modules downstream of C and N assimilation and dissimilation processes and stress responses. Our system was able to discriminate unidentified metabolites and identify bottleneck enzymatic steps. In a comprehensive approach such heuristics become increasingly important because with current technology, the determination of all network components is virtually impossible. For a more precise investigation of biochemical networks, expansion of target metabolites and determination of metabolite levels in each cellular compartment may be suggested. There are technical hurdles, however, in separating organelles without disturbing a wide range of metabolites inside them. Without much technical advancement, therefore, it seems difficult to repeat our time-course measurement for any single cellular compartment although there are reports for such a challenge [46]. Finally, for the analysis part, it is necessary to couple biological information with computer simulations based on large-scale time-resolved measurements of metabolites, proteins, and mRNAs.

Methods

Plant materials

Young seedlings of rice plants, *Oryza sativa* L. ssp. *japonica* Haenuki, at the third-leaf stage were cultured as follows. Rice seeds were germinated on filter paper soaked with Milli-Q water and kept at 30°C in a dark room for 2 days. After germination, the plantlets were placed on rock fiber (35 × 35 × 40 mm; Nittobo, Tokyo, Japan), and grown in a growth chamber (FLI-301N, Tokyo Rika Kikai, Tokyo, Japan) for 18 days. The temperature and light conditions were 25°C and 365 $\mu E \cdot m^{-2}s^{-1}$ for 9 hr (light), 20°C and 0 $\mu E \cdot m^{-2}s^{-1}$ for 11 hr (dark), and 150 $\mu E \cdot m^{-2}s^{-1}$ for 2 hr between light and dark. The plants were watered with Kasugai water culture solution (18.9 mg/L $(NH_4)_2SO_4$, 10.1 mg/L $Na_2HPO_4 \cdot 12H_2O$, 4.7 mg/L KCl, 0.79 mg/L $CaCl_2$, 3.0 mg/L $MgCl_2$, 0.17 mg/L $\cdot FeCl_3 \cdot 6H_2O$, and HCl to adjust the pH to 5.0 – 5.5) [47].

Reagents

Piperazine-1,4-bis(2-ethanesulfonic acid) (PIPES) was purchased from Dojindo (Kumamoto, Japan), methionine sulphone from Avocado Research (Heysham, Lancashire, UK). All other reagents were obtained from conventional commercial sources. Individual stock solutions, at a concentration of 10 or 100 mM, were prepared in Milli-Q water, 0.1 N HCl, or 0.1 N NaOH. The working standard mixture was prepared by diluting these stock solutions with Milli-Q water just before injection. All chemicals used were of analytical or reagent grade. Water

was purified with a Milli-Q purification system (Millipore, Bedford, MA, USA).

Sample preparation

Leaves were harvested (fresh weight approximately 100 mg (6 seedlings)) and frozen in liquid nitrogen to stop enzymatic activity. They were mashed in a Multi-Beads Shocker (Yasuikikai, Osaka, Japan) at 2000 rpm for 10 sec and 0.5 mL of ice-cooled methanol, including 400 μM PIPES and methionine sulphone as an internal standard, was added to dissolve phospholipid membranes and inactive enzymes. Then 0.5 mL ice-cold Milli-Q water was added and the sample was ultrafiltered through a 5-kDa cut-off filter at 9058 g for 10 min to remove proteins, phospholipids, chlorophyll, and other high-molecular-weight impurities. The filtrate was analyzed by CE-MS and CE-DAD methods. To obtain sufficient sensitivity for the analysis of nucleotides, coenzymes, and sugars, the filtrate was concentrated 5-fold by lyophilization [17].

Instruments

All CE-MS experiments were performed by Agilent CE capillary electrophoresis. We used a 1100 series MSD mass spectrometer, a 1100 series isocratic HPLC pump, a G1603A CE-MS adapter kit, and a G1607A CE-ESI-MS sprayer kit (Agilent Technologies). CE-DAD experiments were performed by Agilent CE capillary electrophoresis with a built-in diode-array detector. G2201AA Agilent ChemStation software for CE was used for system control, data acquisition and analysis, and MSD data evaluation.

Analytical conditions

The compounds were analyzed in four groups using three CE-MS methods and one CE-DAD method.

a) Cationic metabolites (amino acids and amines) were analyzed with a fused-silica capillary (50 μm i.d. × 100 cm total length), with 1 M formic acid as the electrolyte. The sample was injected at an injection pressure of 5.0 kPa for 3 sec (approximately 3 nL). The applied voltage was set at 30 kV. The capillary temperature was set to 20°C, and the sample tray was cooled to below 5°C. The sheath liquid (5 mM ammonium acetate in 50% [v/v] methanol-water) was delivered at 10 μL/min. ESI-MS was conducted in positive ion mode; the capillary voltage was set at 4000 V. A flow rate of heated dry nitrogen gas (heater temperature 300°C) was maintained at 10 L/min [12].

b) Anionic metabolites (organic acids and sugar phosphates) were analyzed with a cationic polymer-coated SMILE(+) capillary (Nakalai Tesque, Kyoto, Japan). The electrolyte for CE separation was a 50 mM ammonium acetate solution (pH 8.5). The sample was injected at an injection pressure of 5.0 kPa for 30 sec (approximately 30 nL). The applied voltage was set at -30 kV, and the capillary temperature was set to 30°C. ESI-MS was conducted in negative ion mode; the capillary voltage was set at 3500 V. Other conditions were as in the cationic metabolite analysis [13].

c) Nucleotides and coenzymes were analyzed with an uncharged polymer-coated gas chromatograph capillary, polydimethylsiloxane (DB-1) (Agilent Technologies). The electrolyte for CE separation was 50 mM ammonium acetate solution (pH 7.5). The applied voltage was set at -30 kV and a pressure of 5.0 kPa was added to the inlet capillary during the run. Other conditions were as in the anion analysis [14].

d) Sugars were analyzed with a fused-silica capillary (50 μm i.d. × 112.5 cm total length, 104 cm effective length). Basic anion buffer for CE (Agilent Technologies) was the electrolyte. The sample was injected at a pressure of 5.0 kPa for 10 sec (approximately 10 nL). The applied voltage was set at -25 kV; the capillary temperature, regulated with a thermostat, was 25°C. Sugars were detected by indirect UV detection using a diode-array detector. The signal wavelength was set at 350 nm with a reference at 230 nm [48].

Self-organizing map (SOM) analysis

A free software package, SOM -PAK [49], was used to compute both the SOM and the Sammon map. Before SOM analysis, the observed time-course data for 58 metabolites (including an estimate of S17P) were smoothed by averaging the adjacent data points using a sliding window of width 3, to reduce high-frequency noise presumably originating from individual differences in plant seedlings, rapid oscillations in metabolism, or measurement errors. The missing data points were extrapolated by linear approximation between prior and subsequent data values: Among the 57 metabolites evaluated at 26 time points, only 30 data points could be extrapolated due to the detection limit or contamination of other unidentifiable peaks. The SOM is a map from the input n-dimensional data space (input layer) to a two-dimensional array of nodes (output layer). The vectors in the output layer are the parametric reference vector m_i, which has n elements. An input data vector, x, is compared with m_i, and the best-match vector, which is the smallest Euclidean distance $|x - m_i|$, is mapped onto this location. During learning, nodes that are topographically close in the array up to a certain distance activate each other to learn from the same input vector, and the reference vectors are corrected so that they become close to the input vector. Thus,

$$m_i(t + 1) = m_i(t) + h_{ci}(t)\,[x(t) - m_i(t)],$$

where t is an integer, the discrete-time coordinate, and $h_{ci}(t)$ is the neighborhood kernel, a function defined over

the lattice points. The neighborhood size, N_c, around node c is a function of time, and h_{ci} is defined as

$$h_{ci} = \alpha(t) \quad (i \in N_c)$$
$$h_{ci} = 0 \quad (i \notin N_c),$$

where $\alpha(t)$ is a monotonic decreasing function of time ($0 < \alpha(t) < 1$) called the "learning rate". The learning rate function was defined as

$$\alpha(t) = \alpha(0)(1.0 - t/T),$$

where $\alpha(0)$ is the initial learning rate and T the running length (number of steps) in training. In this study, 58 metabolic time-courses were formatted and classified in a 24 × 24 hexagonal lattice. The applied SOM parameters were: initial radius of the training area = 12, initial learning rate = 0.025, running length = 65 000.

Metabolic pair-wise correlation

Significance levels for Pearson correlation coefficient r were computed depending on the number of metabolite pairs n found throughout the light and dark period, respectively, by calculating t-scores given by $t = r\,(n - 2)^{0.5}/(1 - r)^{0.5}$. The critical t-score was set to correspond to the commonly used p-value of 0.05 in two-sided tests.

Hierarchical clustering

Among several algorithms for clustering analysis, we chose Ward's method [29] in JMP software (ver. 6.0.0; SAS Institute Inc. Cary, NC). Starting from trivial clusters each containing one object only, Ward's method iteratively merges two clusters that will result in the smallest increase in the sum of the square of their differences (i.e., variance). At each step, all possible mergers of two clusters are tried and their variance is computed. The difference between clusters is calculated by the equation:

$$d(a,b) = \frac{n_a n_b}{n_a + n_b}(x_a - x_b)^2$$

Authors' contributions

SS conceived this study, performed the biochemical- and the computational experiments, and wrote the manuscript. MA provided intellectual help for the computational analysis and together wrote the manuscript. TN advised the experimental design. TS and MT supervised the research. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by a grant for the Development of Rice Genome Simulators from MAFF, Japan, and by the Ministry of Education, Culture, Sports, Science and Technology, and a Grant-in-Aid for the 21st Century Center of Excellence (COE) Program entitled "Understanding and Control of Life's Function via Systems Biology (Keio University)". This work was also supported, in part, by Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Fridman E, Pichersky E: **Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products.** *Curr Opin Plant Biol* 2005, **8(3):**242-248.
2. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, Dam K, Oliver SG: **A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.** *Nat Biotechnol* 2001, **19(1):**45-50.
3. Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG, Kell DB: **High-throughput classification of yeast mutants for functional genomics using metabolic footprinting.** *Nat Biotechnol* 2003, **21(6):**692-696.
4. Morgenthal K, Wienkoop S, Scholz M, Selbig J, Weckwerth W: **Correlative GC-TOF-MS-based metabolite profiling and LC-MS-based protein profiling reveal time-related systematic regulation of metabolite-protein networks and improve pattern recognition for multiple biomarker selection.** *Metabolomics* 2005, **1(2):**109-121.
5. Weckwerth W, Loureiro ME, Wenzel K, Fiehn O: **Differential metabolic networks unravel the effects of silent plant phenotypes.** *Proc Natl Acad Sci USA* 2004, **101(20):**7809-7814.
6. Hirai MY, Klein M, Fujikawa Y, Yano Y, Goodenowe DB, Yamazaki Y, Kanaya S, Nakamura Y, Kitayama M, Suzuki H, Sakurai N, Shibata D, Tokuhisa J, Reichelt M, Gershenzon J, Papenbrock J, Saito K: **Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics.** *J Biol Chem* 2005, **280(27):**25590-25595.
7. Kohonen T: *Self-Organizing Maps* Springer-Verlag, Heidelberg, Germany; 1995.
8. Mounet F, Lemaire-Chamley M, Maucourt M, Cabasson C, Giraudel JL, Deborde C, Lessire R, Gallusci P, Bertrand A, Gaudillère M, Rothan C, Rolin D, Moing A: **Quantitative metabolic profiles of tomato flesh and seeds during fruit development: Complementary analysis with ANN and PCA.** *Metabolomics* 2007, **3(3):**273-288.
9. Panagiotou G, Kouskoumvekaki I, Jónsdóttir , Olsson L: **Monitoring novel metabolic pathways using metabolomics and machine learning: Induction of the phosphoketolase pathway in *Aspergillus nidulans* cultivations.** *Metabolomics* 2007, **3(4):**503-516.
10. Fiehn O: **Metabolic networks of *Cucurbita maxima* phloem.** *Phytochemistry* 2003, **62(6):**875-886.
11. Yeung KY, Medvedovic M, Bumgarner RE: **From co-expression to co-regulation: How many microarray experiments do we need?** *Genome Biol* 2004, **5(7):**R48.
12. Soga T, Heiger DN: **Amino acid analysis by capillary electrophoresis electrospray ionization mass spectrometry.** *Anal Chem* 2000, **72:**1236-1241.
13. Soga T, Ueno Y, Naraoka H, Ohashi Y, Tomita M, Nishioka T: **Simultaneous determination of anionic intermediates for *Bacillus subtilis* metabolic pathway by capillary electrophoresis electrospray ionization mass spectrometry.** *Anal Chem* 2002, **74:**2233-2239.
14. Soga T, Ueno Y, Naraoka H, Matsuda K, Tomita M, Nishioka T: **Pressure-assisted capillary electrophoresis electrospray ionization mass spectrometry for analysis of multivalent anions.** *Anal Chem* 2002, **74:**6224-6229.
15. Soga T, Ohashi Y, Ueno Y, Naraoka H, Tomita M, Nishioka T: **Quantitative metabolome analysis using capillary electrophoresis mass spectrometry.** *J Proteome Res* 2003, **2:**488-494.
16. Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, Hirasawa T, Naba M, Hirai K, Hoque A, Ho PY, Kakazu Y, Sugawara K, Igarashi S, Harada S, Masuda T, Sugiyama N, Togashi T, Hasegawa M, Takai Y, Yugi K, Arakawa K, Iwata N, Toya Y, Nakayama Y, Nishioka T, Shimizu K, Mori H, Tomita M: **Multiple high-throughput analyses monitor the response of E. coli to perturbations.** *Science* 2007, **316(5824):**593-7.
17. Sato S, Soga T, Nishioka T, Tomita M: **Simultaneous determination of the main metabolites in rice leaves using capillary**

electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *The Plant J* 2004, **40**:151-163.
18. Sammon JW Jr: **A nonlinear mapping for data structure analysis.** *IEEE Transactions Computers* 1969, **C-18(5)**:401-409.
19. **KEGG pathway database** [http://www.genome.ad.jp/kegg/pathway.html]
20. **Swiss-Prot database** [http://au.expasy.org/sprot/]
21. **Rice Annotation Project Data Base** [http://rapdb.dna.affrc.go.jp/]
22. Stafford HA, Magaldi A, Vennesland B: **The enzymatic reduction of hydroxypyruvic acid to D-glyceric acid in higher plants.** *J Biol Chem* 1954, **207**:621-629.
23. Rippert P, Matringe M: **Purification and kinetic analysis of the two recombinant arogenate dehydrogenase isoforms of *Arabidopsis thaliana*.** *Eur J Biochem* 2002, **269**:4753-4761.
24. Boldt R, Edner C, Kolukisaoglu U, Hagemann M, Weckwerth W, Wienkoop S, Morgenthal K, Bauwe H: **D-Glycerate 3-kinase, the last unknown enzyme in the photorespiratory cycle in Arabidopsis, belongs to a novel kinase family.** *The Plant Cell* 2005, **17**:2413-2420.
25. Duncan K, Edwards RM, Coggins JR: **The pentafunctional *arom* enzyme of *Saccharomyces cervisiae* is a monofunctional domains.** *Biochem J* 1987, **246**:375-386.
26. Kuhr WG: **Separation of small organic molecules.** In *Capillary Electrophoresis: Theory and Practice* Edited by: Camilleri P. CRC Press LLC, Boca Raton, FL, USA; 1997:91-133.
27. **KEGG ligand database** [http://www.genome.ad.jp/kegg/ligand.html]
28. Steuer R, Kurths J, Fiehn O, Weckwerth W: **Observing and interpreting correlations in metabolic networks.** *Bioinformatics* 2003, **19(8)**:1019-1026.
29. Ward JH: **Hierarchical grouping to optimize an objective function.** *J Am Stat Assoc* 1963, **58**:236-245.
30. Ashihara H, Sato F: **Pyrophosphate: Fructose-6-phosphate 1-phosphotransferase and biosynthetic capacity during differentiation of hypocotyls of Vigna seedlings.** *Biochim Biophys Acta* 1993, **1156**:123-127.
31. Farr TJ, Huppe HC, Turpin DH: **Coordination of chloroplastic metabolism in N-limited chlamydomonas reinhardtii by redox modulation (I. The activation of phosphoribulokinase and glucose-6-phosphate dehydrogenase is relative to the photosynthetic supply of electrons).** *Plant Physiol* 1994, **105**:1037-1042.
32. Woodrow IE, Berry J: **Enzymatic regulation of photosynthetic CO_2 fixation in C_3 plants.** *Ann Rev Plant Physiol Plant Molec Biol* 1988, **39**:533-594.
33. Lea PJ, Ireland RJ: **Nitrogen metabolism in higher plants.** In *Plant Amino Acids, Biochemistry and Biotechnology* Edited by: Singh BK. New York: Marcel Dekker, Inc; 1999:1-47.
34. Ireland RJ, Lea PJ: **The enzymes of glutamine, glutamate, asparagines, and aspartate metabolism.** In *Plant Amino Acids, Biochemistry and Biotechnology* Edited by: Singh BK. New York: Marcel Dekker, Inc; 1999:49-109.
35. Scheible WR, Krapp A, Stitt M: **Reciprocal diurnal changes of phosphoenol pyruvate carboxylase expression and cytosolic pyruvate kinase, citrate synthase and NADP-isocitrate dehydrogenase expression regulate organic acid metabolism during nitrate assimilation in tobacco leaves.** *Plant Cell and Environ* 2000, **23**:1155-1167.
36. Noctor G, Novitskaya L, Lea PJ, Foyer CH: **Co-ordination of leaf minor amino acid contents in crop species: Significance and interpretation.** *J Exp Bot* 2002, **53(370)**:939-945.
37. Ferrario-Méry S, Suzuki A, Kunz C, Valadier MH, Roux Y, Hirel B, Foyer CH: **Modulation of amino acid metabolism in transformed tobacco plants deficient in Fd-GOGAT.** *Plant and Soil* 2000, **221**:67-79.
38. Rao IM, Arulanantham AR, Terry N: **Diurnal changes in adenylates and nicotinamide nucleotides in sugar beet leaves.** *Photosynthesis Res* 1990, **23**:205-212.
39. Bonzon M, Hug M, Wagner E, Greppin H: **Adenine nucleotides and energy charge evolution during the induction of flowering in spinach leaves.** *Planta* 1981, **152**:189-194.
40. Stitt M, Lilley RM, Heldt HW: **Adenine nucleotide levels in the cytosol, chloroplasts, and mitochondria of wheat leaf protplasts.** *Plant Physiol* 1982, **70**:971-977.
41. Chen LS, Nose A: **Day-Night changes of energy-rich compounds in crassulacean acid metabolism (CAM) species utilizing hexose and starch.** *Ann Bot* 2004, **94**:449-455.
42. Leegood RC: **Photosynthesis in C_3 plants: The Benson-Calvin cycle and photorespiration.** In *Plant Biochemistry and Molecular Biology* 2nd edition. Edited by: Lea PJ, Leegood RC. John Wiley & Sons Ltd; 1999:29-50.
43. Hampp R, Schnabl H: **Adenine and pyridine nucleotide status of isolated Vicia guard cell protoplasts during K+-induced swelling.** *Plant and Cell Physiol* 1984, **25(7)**:1233-1239.
44. May MJ, Vernoux T, Leaver C, Montagu MV, Inze D: **Glutathione homeostasis in plants: Implications for environmental sensing and plant development.** *J Exp Bot* 1998, **49(321)**:649-667.
45. Lovaas E: **Antioxidant and metal-chelating effects of polyamines.** In *Advances in Pharmacology. Antioxidants in Disease Mechanisms and Therapy 38* Edited by: Sies H. New York: Academic Press; 1996:119-149.
46. Farré EM, Tiessen A, Roessner U, Geigenberger P, Trethewey RN, Willmitzer L: **Analysis of the compartmentation of glycolytic intermediates, nucleotides, sugars, organic acids, amino acids, and sugar alcohols in potato tubers using a nonaqueous fractionation method.** *Plant Physiol* 2001, **127**:685-700.
47. Kasugai S: **Studies of water culture.** *Jpn J Soil Sci Plant Nutr* 1939, **13**:669-822. (in Japanese)
48. Soga T, Ross GA: **Simultaneous determination of inorganic anions, organic acids, amino acids and carbohydrate by capillary electrophoresis.** *J Chromatogr A* 1999, **837**:231-239.
49. **SOM-PAK** [http://www.cis.hut.fi/research/som_lvq_pak]

ORIGINAL PAPER

Vol. 24 no. 14 2008, pages 1590–1595
doi:10.1093/bioinformatics/btn240

Gene expression

Aligning LC peaks by converting gradient retention times to retention index of peptides in proteomic experiments

Kosaku Shinoda[1,2], Masaru Tomita[1,2] and Yasushi Ishihama[1,3,*]

[1]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, [2]Human Metabolome Technologies, Inc., Tsuruoka, Yamagata 997-0052 and [3]PRESTO, Japan Science and Technology Agency, Sanbancho Bldg., 5-Sanbancho, Chiyodaku, Tokyo 102-0075, Japan

Received on February 14, 2008; revised on April 29, 2008; accepted on May 17, 2008

Advance Access publication May 19, 2008

Associate Editor: David Rocke

ABSTRACT

Motivation: Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is a powerful tool in proteomics studies, but when peptide retention information is used for identification purposes, it remains challenging to compare multiple LC-MS/MS runs or to match observed and predicted retention times, because small changes of LC conditions unavoidably lead to variability in retention times. In addition, non-contiguous retention data obtained with different LC-MS instruments or in different laboratories must be aligned to confirm and utilize rapidly accumulating published proteomics data.

Results: We have developed a new alignment method for peptide retention times based on linear solvent strength (LSS) theory. We found that log k_0 (logarithm of retention factor for a given organic solvent) in the LSS theory can be utilized as a 'universal' retention index of peptides (RIP) that is independent of LC gradients, and depends solely on the constituents of the mobile phase and the stationary phases. We introduced a machine learning-based scheme to optimize the conversion function of gradient retention times (t_g) to log k_0. Using the optimized function, t_g values obtained with different LC-MS systems can be directly compared with each other on the RIP scale. In an examination of *Arabidopsis* proteomic data, the vast majority of retention time variability was removed, and five datasets obtained with various LC-MS systems were successfully aligned on the RIP scale.

Contact: y-ishi@ttck.keio.ac.jp

1 INTRODUCTION

Liquid chromatography-mass spectrometry (LC-MS) is a powerful tool for the separation and identification of peptides in proteomics studies. While several methods and software tools are available for identifying peptides/proteins from mass spectra, the high complexity of a digested proteome and the vastly larger number of possible peptide sequences make accurate peptide/protein identification challenging. As the chromatographic retention times of peptides depend on their amino acid sequences, their retention times complement the information provided by MS and thus enhance their identifiability (Palmblad *et al.*, 2002; Petritis *et al.*, 2003).

Comparing multiple LC-MS/MS runs or matching observed and predicted retention times for identification purposes remains a challenging issue, because small changes in flow rate, column length, column packing, void volume and mobile phase composition unavoidably lead to variability in retention times. In addition, it was recently reported that even changing pore size of chromatographic beads as well as the ion-pair reagents such as trifluoroacetic acid, heptafluorobutyric acid and acetic acid in the mobile phase affects the peptide retention times significantly (Ishihama *et al.*, 2008; Krokhin, 2006). Furthermore, non-contiguous retention data obtained with different LC-MS instruments or in different laboratories must be aligned to confirm and utilize published proteomics data.

A widely used approach to the chromatographic-alignment problem is to fit a piecewise linear function to maximize the correlation between the samples. Methods of this kind are often characterized as correlation optimized warping (COW) (Nielsen *et al.*, 1998), and several derivative methods have been investigated (van Nederkassel *et al.*, 2006). In principle, this approach can be extended to aligning multi-dimensional data. However, the handling of proteomics data is extremely difficult because the data are typically characterized by a very large input dimension (i.e. tryptic peptides). Thus, more sophisticated alignment algorithms are needed to extract higher quality information from large-scale LC-MS-based experiments.

Several approaches for the alignment of peptide retention times have been developed and applied to high-throughput proteomics. For example, in the accurate mass and time tag (AMT) approach (Callister *et al.*, 2006; Jaitly *et al.*, 2006; Norbeck *et al.*, 2005; Smith *et al.*, 2002; Zimmer *et al.*, 2006), results from different LC-MS or MS/MS datasets are combined by finding the conversion functions of mass and retention times that are required to remove variability in mass and retention time measurements between analyses. Machine learning has also been applied to develop an 'intelligent' system for comparing large numbers of LC/MS experiments. The genetic algorithm (GA) has enabled the optimization of two variables of the linear normalization function for each LC separation so as to reduce the variance function of specific peptides, i.e. the regressed retention times for each separation (Petritis *et al.*, 2003). While this approach has generated excellent results, the normalization approach becomes time-prohibitive as the number of peptides used increases significantly, due to the many generations (iterations) required to

*To whom correspondence should be addressed.

align all analyses (Petritis *et al.*, 2006). To remove this limitation, Strittmatter *et al.* (2003) regressed, observed retention times of confidently identified peptides to predicted normalized elution time (NET) of the sequences using a quadratic function for each LC-MS run. The obtained quadratic equations were used to convert observed retention times to observed NET, and all LC-MS runs could be compared on scales of the NET. However, due to their use of an in-house-built nanoflow pump with ultrahigh pressure tolerance, it would be difficult to apply their NET scale to other datasets obtained with commercial systems in other proteomics laboratories, because their nanoflow pump generates exponential gradient curves depending on the flow-rate (Shen *et al.*, 2001).

Here we report the development of a new alignment method using $\log k_0$ (logarithm of retention factor for a given organic solvent) from linear solvent strength (LSS) theory (Stadalius *et al.*, 1984). Peptide LC-MS data are aligned by converting different gradient retention time scales to a single scale of predicted $\log k_0$. We introduce a GA to optimize the conversion function between retention times and $\log k_0$. Using the optimized function, peptide retention times obtained from different gradients and/or LC-MS systems can be compared with each other on the same $\log k_0$ scale. Unlike other functional optimization-based alignment techniques, realignments after each new experiment are not required, and thus the technical weaknesses of GA are overcome. The new method was applied to the soluble fraction of *Arabidopsis* cells and datasets obtained with various LC-MS systems were successfully aligned.

2 MATERIALS AND METHODS

2.1 Preparation of cell lysates

Escherichia coli MC4100 cells (see Section 3.1) were grown at 37°C in rich medium as described (Kerner *et al.*, 2005), and were lyzed by ultrasonication and centrifuged at 3000 × *g* for 10 min to collect the supernatants. *Arabidopsis* (ecotype Landsberg erecta) cells were a generous gift from Dr H. Nakagami (Riken, Yokohama, Japan). The frozen cells were disrupted with a Multi-beads shocker (MB400U, Yasui Kikai, Tokyo, Japan) and suspended in 0.1 M Tris–HCl (pH 8.0). The supernatants were collected by centrifugation at 1500*g* for 10 min.

2.2 Sample preparation

Proteins from these cell lysates were dried and resuspended in 50 mM Tris–HCl buffer (pH 9.0) containing 8 M urea. The mixtures were individually reduced with dithiothreitol (DTT), alkylated with iodoacetamide and digested with Lys-C, followed by dilution and trypsin digestion as described (Saito *et al.*, 2006). The digested samples were then desalted using StageTips with C18 Empore disk membranes (Rappsilber *et al.*, 2007).

2.3 NanoLC-MS/MS analysis

All samples were analyzed by nanoLC-MS/MS using a QSTAR Pulsar i mass spectrometer (AB/MDS-Sciex, Toronto, Canada) equipped with an Agilent 1100 nanoflow pump (Waldbron, Germany) or an LTQ-Orbitrap mass spectrometer (Thermofisher, Bremen, Germany) with a Dionex Ultimate 300 pump. In both systems, an HTC-PAL autosampler (CTC Analytics AG, Zwingen, Switzerland) equipped with a Valco C2 valve with 150 μm ports as an injection valve was used. ReproSil-Pur 120 C18-AQ materials (3 μm, Dr Maisch, Ammerbuch, Germany) were packed into a self-pulled needle (100 μm ID, 6 μm opening, 150 mm length) with a nitrogen-pressurized column loader cell (Nikkyo Technos, Tokyo, Japan) to prepare an analytical column needle with 'stone-arch' frit (Ishihama *et al.*, 2002). A spray voltage of 2400 V was applied via the metal connector as described (Ishihama *et al.*, 2002). The injection volume was 5 ml and the flow rate was 500 nl/min. The mobile phases consisted of (A) 0.5% acetic acid in water and (B) 0.5% acetic acid in 80% acetonitrile. Four linear gradient conditions of 5% B to 60% in 30, 60, 120 and 180 min were employed. Four MS/MS scans (0.6 s each) per one MS scan (1 s) were performed with the QSTAR, whereas the top 10 precursors were selected for MS/MS scans for the LTQ-Orbitrap. The scan range was m/z 350–1400 for the QSTAR and 300–1500 for the LTQ-Orbitrap.

2.4 Data analysis

MS peak lists were created by scripts in Analyst QS (MDS-Sciex) on the basis of the recorded fragmentation spectra, and were submitted to the Mascot database search engine (Matrix Science, London, UK) against the SwissProt database (release 45.0) to identify proteins from *E.coli* samples, while the TAIR version 7 (April 25, 2007) database was used for *Arabidopsis* samples. The following search parameters were used in all Mascot searches: maximum of two missed trypsin cleavages, cycteine carbamidomethylation as a fixed modification and methionine oxidation as a variable modification. A precursor mass tolerance of 0.2 Da and a fragment ion mass tolerance of 0.2 Da were set for the QSTAR, whereas a precursor mass tolerance of 3 p.p.m. and a fragment ion mass tolerance of 0.8 Da were used for the LTQ-Orbitrap. All peptides with scores less than the identity threshold ($P \geq 0.05$) or a rank >1 were automatically discarded.

2.5 Measurement of retention factors from gradient analysis

The reversed-phase retention factor k is generally described as

$$\log k = \log k_0 - S\phi \tag{1}$$

where ϕ is the volume fraction of the less polar component in the water–organic mobile phase, k_0 is the value of k for the solute at the start of the gradient in the initial mobile phase ($\phi = 0$) and S is a constant characteristic for a given analyte and chromatographic system (Stadalius *et al.*, 1984).

Solute retention time t_g in gradient elution is given as

$$t_g = \left(\frac{t_0}{b}\right)\left[\log 2.3 k_0 b\left(\frac{t_{scc}}{t_0}\right)+1\right]+t_{scc}+t_D \tag{2}$$

where t_0 is the column dead-time for a small solute molecule, t_{scc} is the value of t_0 for the solute in question, t_D is the dwell-time of the gradient system and b is a gradient parameter defined by

$$b = S\Delta\phi t_0/t_G \tag{3}$$

Here the quantity t_G is the gradient time and $\Delta\phi$ is the change in ϕ during the gradient ($\Delta\phi = 1$ for a 0–100% gradient) (Snyder, 1980). For smaller solutes and larger pore particles, Equation (2) can be approximated by

$$t_g = \left[\frac{t_G}{(S\Delta\phi)}\right]\log\left[2.3 k_0 t_0\left(\frac{S\Delta\phi}{t_G}\right)+1\right]+t_0+t_D \tag{4}$$

By solving Equation (4) for k_0, Equation (5) is derived:

$$k_0 = \frac{t_G\left(-1+10^{-t_G/\Delta\phi S\left(t_0+t_D-t_g\right)}\right)}{\Delta\phi 2.3 S t_0} \tag{5}$$

Four gradient elution runs were performed for *E.coli* samples as described above, and the observed t_g, $t_G/\Delta\phi$, t_0, t_D values were substituted into Equation (4). A Microsoft Excel multi-line fitting program based on the semi-Newton method was run to optimize S and k_0 values in order to minimize the sum of the differences between calculated and observed t_g values. The obtained S and k_0 were used as observed values for further analysis.

Table 1. Experimental parameters for the genetic algorithm used

Experimental parameters	Parameter value
Number of maximum generation G	200
Number of individual P	500
Crossover ratio c (%)	45
Crossover strategy	Uniform
Selection strategy	Roulette
Mutation ratio m (%)	45

2.6 Implementation of the algorithm

Obtained k_0 was used to construct the $\log k_0$ predictor. We employed a three-layer artificial neural network (ANN) with back-propagation learning. A sigmoid function was applied to each node in the ANN. To reduce unnecessarily large parameters (weights) among nodes, the pruning method was used as described (Shinoda *et al.*, 2006). The ANN software used was JMP software, version 6.0.2 (SAS Institute, Cary, NC, USA). Experimental retention time (t_g) was converted to predicted $\log k_0$ using Equation (5) containing several parameters ($t_G/\Delta\phi$, t_0, t_D, S). Among them, S was predicted for each identified peptide using a previously reported ANN based on the dependence of S on the amino acid composition (Ishihama, 2006), and the remaining parameters ($t_G/\Delta\phi$, t_0, t_D) were optimized using GA. Our GA was implemented in Perl language with the AI::Genetic module from CPAN (www.cpan.org). The numerical experimental conditions are shown in Table 1. These computational portions of our work were performed on a Pentium 4 Xeon 2.0 GHz CPU.

3 RESULTS AND DISCUSSION

3.1 Prediction of $\log k_0$ using an ANN

We analyzed *E.coli* samples under four different linear gradient conditions and obtained the data pairs of $\log k_0$ and S for 278 peptides. The correlation coefficients between observed and calculated t_g values per each peptide ranged from 0.9993 to 1.000 for four data points from 30 min to 180 min gradient runs, indicating that LSS theory was valid for the peptides in this range. In order to predict $\log k_0$ values from peptide sequences, we trained an ANN using the number of residues of each amino acid in the identified *E.coli* peptides as inputs and obtained $\log k_0$ as outputs based on the assumption that $\log k_0$ of peptides depends on amino acid composition. We adopted three-layer architecture for the ANN because it could approximate any function (Funahashi, 1989). We tried hidden nodes ranging from 2 to 10, and the $\log k_0$ response curves of each input variable, constituting an approximate function from sampled values, were used as the criteria for determining the number of hidden nodes. We added hidden nodes until the response curves were not too flexible or non-linear. Consequently, we adopted five nodes in the hidden layer and our ANN had a 20-5-1 architecture. Other parameters for ANN training (training ratio, momentum and random numbers for initial ANN weights) were determined empirically. Each of the trainings was continued until the epoch (iteration) reached 100 or until improvement of the optimization function fell below a learning convergence criterion. Figure 1 is a global comparison between predicted and measured $\log k_0$ for 278 peptides through 10-fold two-deep cross-validations (Jonathan *et al.*, 2000). Overall, our results were satisfactory; the coefficient of determination (R^2) was 0.8895 and the mean prediction error

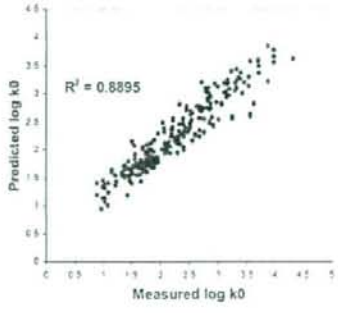

Fig. 1. The correlation between experimentally measured and predicted $\log k_0$ for all peptides derived from *E.coli* K12 proteome through 10-fold two-deep cross-validations.

was 0.189±7.1% (relative standard deviation, RSD). These results support the validity of our assumption that the $\log k_0$ of peptides depends on amino acid composition. ANNs have recently been utilized for accurate modeling of peptide retention time (Petritis *et al.*, 2003, 2006; Shinoda *et al.*, 2006), but application to $\log k_0$ prediction has not yet been reported. We used this ANN predictor for the following GA-based optimization of the conversion function.

The scheme of our alignment approach is illustrated in Figure 2. S values of identified peptides were computationally predicted using a previously reported ANN (Ishihama, 2006) from amino acid composition. The ANN predictor eliminated the need for multiple chromatographic runs for derivations of S and enabled experimental $\log k_0$ to be obtained from a single LC-MS run. On the other hand, the constructed ANN enabled predicted $\log k_0$ to be obtained from the amino acid composition of peptides. The conversion function [Equation (5)] was optimized with a GA using the sum of squared errors (SSE) function between experimental and predicted $\log k_0$ as an evaluation function. Optionally, we adjusted GA-optimized $\log k_0$ values using the linear relationship, if necessary. This conversion enabled various LC gradient data to be compared on the same scale of RIP. RIP is a converted $\log k_0$ scale on a time scale of the $\log k_0$ predictor, which is specific for a given set of gradient analyses with a given mobile phase and columns, i.e. the *E.coli* dataset in this article. Using the optimized function, peptide retention times obtained from different LC-MS systems and/or gradients can be directly compared on the same RIP dimension and easily aligned.

3.2 Application to *Arabidopsis* proteome data

To demonstrate the usability of our alignment algorithm, we conducted an independent validation study with real complex samples (*Arabidopsis* cells). The proteomics sample was prepared according to the above protocol and analyzed using two different LC-MS systems under five different LC conditions (Table 2). Peptides were identified for each LC-MS run using Mascot. The number of identified peptides was 605, 1050, 980, 3861 and 5719 for conditions 1–5, respectively. Experimental t_g was converted to RIP using Equation (5). Parameters were optimized using the GA so that the difference between predicted and converted RIP of identified peptides was minimized. We used the GA because it

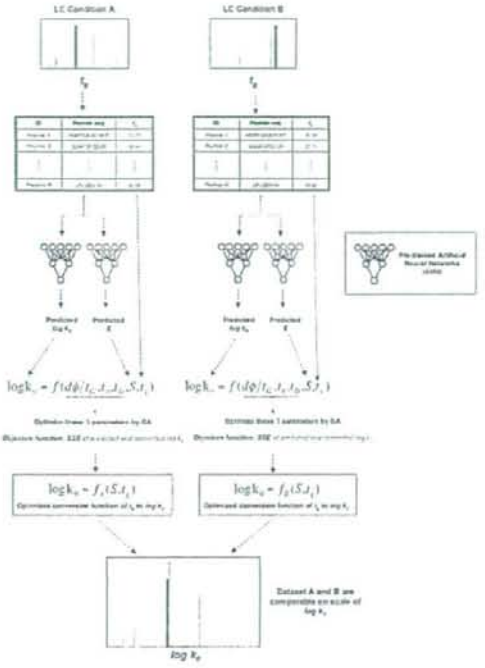

Fig. 2. Schematic flowchart depicting the method. In this example, peptide retention data (t_g) obtained with two different LC-MS systems (A and B) are aligned. S and $\log k_0$ of identified peptides are computationally predicted using a pretrained ANN based on amino acid composition determined by MS/MS ion search (e.g. Mascot). Parameters of the conversion functions of t_g to $\log k_0$ are optimized for each LC condition based on the predicted S and predicted $\log k_0$ values using a GA. The objective function is the SSE between predicted and converted (experimental) $\log k_0$. After functional optimization, datasets A and B are comparable on the same $\log k_0$ (RIP) scale. This algorithm is easily expandable to three or more samples.

Table 2. LC system and gradients used for method validation

Condition	LC-MS systems	Gradient (min)	Column
1	Agilent1100-QSTAR	30	Column 1 (100 μm ID/8 cm L)
2	Agilent1100-QSTAR	60	Column 1 (100 μm ID/8 cm L)
3	Agilent1100-QSTAR	60	Column 2 (100 μm ID/15 cm L)
4	Ultimate3000-Orbitrap	60	Column 2 (100 μm ID/15 cm L)
5	Ultimate3000-Orbitrap	120	Column 2 (100 μm ID/15 cm L)

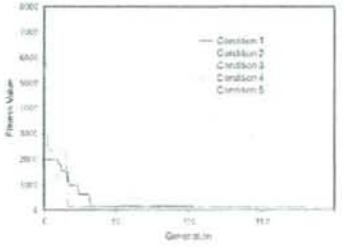

Fig. 3. Changes in the fitness values of best-of-generation individuals for each experimental condition (1–5).

can determine many parameters simultaneously with high accuracy, and selected the real-coded GA (Janikow and Michalewicz, 1991) because it improves the optimization speed compared with the conventional binary GA. The time required for one trial was ~1 h. The experiments were conducted in 50 trials with different random seeds. Comparison of the trajectories shows that fitness values decreased until ~60 generations (Fig. 3). The R^2 between the predicted and experimental $\log k_0$ was 0.9604–0.9968. These results indicate the value of GA in functional optimization for gradient retention time conversion. Unlike traditional non-linear regression, GA-based approaches offer advantages that include a capacity to self-learn and to obtain optimized parameters without the need for time-consuming manual tunings and detailed understanding of the characteristics of functions.

The results of conversion using the optimized function are shown in Figure 4. The converted $\log k_0$ (RIP) of peptides identified among the different LC conditions are plotted. On the RIP scale, most peptides are on the locus of $y=x$ (Spearman $r=0.9863$–0.9988) despite the difference of columns (**A**), systems (**B**) and gradients (**C**). RIP was still effective where columns, systems and gradients were all different (**D**). Using RIP, retention of commonly identified peptides can be compared on the same scale and we can easily validate proteomic data across various LC-MS systems. Our method is more effective when three or more different LC-MS datasets should be aligned. RIP is a general parameter, and thus reoptimization is not required even when a new dataset for comparison is added.

3.3 Probability scores and Δ RIP

As RIP depends on the amino acid sequence, comparison of predicted and experimental (converted) RIP allows validation of peptide sequences determined by MS/MS ion search, i.e. peptides which have ΔRIP above a certain level are more likely to be false positives. The relationship between Mascot probability score, which indicates reliability of peptide identification, and ΔRIP for *Arabidopsis* data is shown in Figure 5. This showed a negative correlation between ΔRIP and score. Peptides with low reliability (probability score <16) have a larger proportion of 'outlier' peptides, while ΔRIP of a majority of reliable (>95%) peptides is less than 0.5. This indicates the validity of the converted RIP and our predictors. Among the reliable peptides, the threshold value of a 5% outlier in ΔRIP was 0.552. This result indicates that peptide

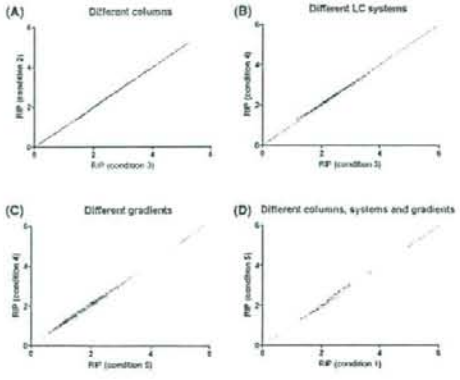

Fig. 4. Correlation of converted RIP among commonly identified peptides between experiments with different columns, LC systems and/or gradients. Black slant line indicates $y = x$.

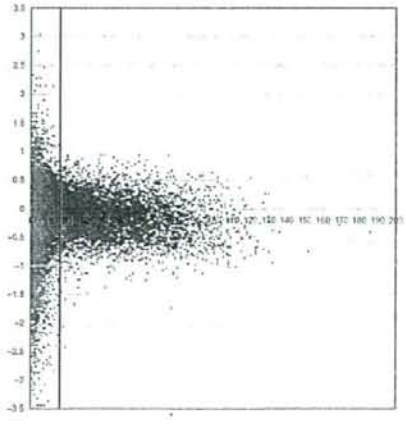

Fig. 5. Relationship between ΔRIP (predicted–experimental) and Mascot probability score. The result for condition 5 (Table 2) is shown. The bold vertical line indicates probability score 16 (>16 scores indicate >95% reliability).

identification where ΔRIP is more than 0.55 is very likely to be a misidentification.

4 CONCLUSION

We have developed a new alignment method for LC-MS-based proteomics data using GA-based optimization of the conversion function between gradient retention times and the logarithm of retention factor ($\log k_0$). The method was applied to the soluble fraction of *Arabidopsis* cells, and five datasets obtained with different LC gradients were appropriately aligned. Converted $\log k_0$ (RIP) values can be used between laboratories as long as the stationary phase and the mobile phase are identical. This method should be useful for comparing proteomics datasets between laboratories and for utilizing the rapidly accumulating published proteomics LC-MS data. In addition, this method is also applicable for peptide mixtures containing partially modified amino acid residues such as phosphorylated serine, threonine and tyrosine. Since the post-translational modifications (PTM) such as phosphorylation are quite important to understand cellular functions, this method would be helpful to perform PTM proteome analysis. Further studies are in progress in our laboratory.

ACKNOWLEDGEMENTS

We thank Yasuyuki Igarashi and Mikiko Hattori (Keio University) for their technical support.

Funding: This work was supported by research funds from the Yamagata prefectural government and Tsuruoka city.

Conflict of Interest: none declared.

REFERENCES

Callister,S.J. *et al.* (2006) Application of the accurate mass and time tag approach to the proteome analysis of sub-cellular fractions obtained from Rhodobacter sphaeroides 2.4.1. Aerobic and photosynthetic cell cultures. *J. Proteome Res.*, **5**, 1940–1947.

Funahashi,K. (1989) On the approximate realization of continuous mappings by neural networks. *Neural Netw.*, **2**, 183–192.

Ishihama,Y. (2006) Method for detection of peptide sequence based on chromatography retention time, PCT/JP2006/315549.

Ishihama,Y. *et al.* (2002) Microcolumns with self-assembled particle frits for proteomics. *J. Chromatogr. A*, **979**, 233–239.

Ishihama,Y. *et al.* (2008) Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics*, **9**, 102.

Jaitly,N. *et al.* (2006) Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.*, **78**, 7397–7409.

Janikow,C.Z. and Michalewicz,Z. (1991) An experimental comparison of binary and floating point representations in genetic algorithms. In *Proceedings of the Fourth International Conference on Genetic Algorithms*. Morgan Kaufmann, San Diego, CA, USA, pp. 31–36.

Jonathan,P. *et al.* (2000) On the use of cross-validation to assess performance in multivariate prediction. *Stat. Comput.*, **10**, 209–229.

Kerner,M.J. *et al.* (2005) Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli. *Cell*, **122**, 209–220.

Krokhin,O.V. (2006) Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal. Chem.*, **78**, 7785–7795.

Nielsen,N.-P.V. *et al.* (1998) Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A*, **805**, 17–35.

Norbeck,A.D. *et al.* (2005) The utility of accurate mass and LC elution time information in the analysis of complex proteomes. *J. Am. Soc. Mass Spectrom.*, **16**, 1239–1249.

Palmblad,M. *et al.* (2002) Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal. Chem.*, **74**, 5826–5830.

Petritis,K. *et al.* (2003) Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.*, **75**, 1039–1048.

Petritis,K. *et al.* (2006) Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.*, **78**, 5026–5039.

Rappsilber,J. *et al.* (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoco.*, **2**, 1896–1906.

Saito,H. *et al.* (2006) Multiplexed two-dimensional liquid chromatography for MALDI and nanoelectrospray ionization mass spectrometry in proteomics. *J. Proteome Res.*, **5**, 1803–1807.

Shen,Y. *et al.* (2001) Packed capillary reversed-phase liquid chromatography with high-performance electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. *Anal. Chem.*, **73**, 1766–1775.

Shinoda,K. *et al.* (2006) Prediction of liquid chromatographic retention times of peptides generated by protease digestion of the Escherichia coli proteome using artificial neural networks. *J. Proteome Res.*, **5**, 3312–3317.

Smith,R.D. *et al.* (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, **2**, 513–523.

Snyder,L.R. (1980) *High Performance Liquid Chromatography: Advances and Perspectives*. Academic Press, New York.

Stadalius,M.A. *et al.* (1984) Optimization model for the gradient elution separation of peptide mixtures by reversed-phase high-performance liquid chromatography: verification of retention relationships. *J. Chromatogr. A*, **296**, 31–59.

Strittmatter,E.F. *et al.* (2003) Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **14**, 980–991.

van Nederkassel,A.M. *et al.* (2006) A comparison of three algorithms for chromatograms alignment. *J. Chromatogr. A*, **1118**, 199–210.

Zimmer,J.S. *et al.*, (2006) Advances in proteomics data analysis and display using an accurate mass and time tag approach. *Mass Spectrom. Rev.*, **25**, 450–482.

Construction of a Biological Tissue Model Based on a Single-Cell Model: A Computer Simulation of Metabolic Heterogeneity in the Liver Lobule

Hiroshi Ohno**,†
Keio University

Yasuhiro Naito*,**,†,‡
Keio University

Hiromu Nakajima§
Osaka Medical Center
for Cancer and
Cardiovascular Diseases

Masaru Tomita**,†,‡
Keio University

Abstract An enormous body of information has been obtained by molecular and cellular biology in the last half century. However, even these powerful approaches are not adequate when it comes to higher-level biological structures, such as tissues, organs, and individual organisms, because of the complexities involved. Thus, accumulation of data at the higher levels supports and broadens the context for that obtained on the molecular and cellular levels. Under such auspices, an attempt to elucidate mesoscopic and macroscopic subjects based on plentiful nanoscopic and microscopic data is of great potential value. On the other hand, fully realistic simulation is impracticable because of the extensive cost entailed and enormous amount of data required. Abstraction and modeling that balance the dual requirements of prediction accuracy and manageable calculation cost are of great importance for systems biology. We have constructed an ammonia metabolism model of the hepatic lobule, a histological component of the liver, based on a single-hepatocyte model that consists of the biochemical kinetics of enzymes and transporters. To bring the calculation cost within reason, the porto-central axis, which is an elemental structure of the lobule, is defined as the systems biological unit of the liver, and is accordingly modeled. A model including both histological structure and position-specific gene expression of major enzymes largely represents the physiological dynamics of the hepatic lobule in nature. In addition, heterogeneous gene expression is suggested to have evolved to optimize the energy efficiency of ammonia detoxification at the macroscopic level, implying that approaches like this may elucidate how properties at the molecular and cellular levels, such as regulated gene expression, modify higher-level phenomena of multicellular tissue, organs, and organisms.

Keywords
Zonal metabolic heterogeneity, hepatic lobule, biological simulation, ammonia metabolism

* Contact author.
** Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, Tsuruoka, 997-0035, Japan. E-mail: n02139ho@sfc.keio.ac.jp
† Bioinformatics Program, Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, 252-8520, Japan.
‡ Department of Environmental Information, Keio University, 5322 Endo, Fujisawa, 252-8520, Japan. E-mail: ynaito@sfc.keio.ac.jp (Y.N.); mt@sfc.keio.ac.jp (M.T.)
§ Clinical Laboratory, Osaka Medical Center for Cancer and Cardiovascular Diseases, 1-3-3 Nakamichi, Higashinariku, Osaka, 537-0025, Japan. E-mail: nakajima-hi@mc.pref.osaka.jp