Metabolomics Vol. 1, No. 2, April 2005 (© 2005)
DOI: 10.1007/s11306-005-4430-9

Correlative GC-TOF-MS-based metabolite profiling and LC-MS-based protein profiling reveal time-related systemic regulation of metabolite–protein networks and improve pattern recognition for multiple biomarker selection

Katja Morgenthal, Stefanie Wienkoop, Matthias Scholz, Joachim Selbig, and Wolfram Weckwerth*

Max Planck Institute of Molecular Plant Physiology, 14424, Potsdam, Germany

Received 14 January 2005; accepted 11 March 2005

A novel approach is presented combining quantitative metabolite and protein data and multivariate statistics for the analysis of time-related regulatory effects of plant metabolism at a systems level. For the analysis of metabolites, gas chromatography coupled to a time-of-flight mass analyzer (GC-TOF-MS) was used. Proteins were identified and quantified using a novel procedure based on shotgun sequencing as described recently (Weckwerth *et al.*, 2004b, *Proteomics* **4**, 78–83). For comparison, leaves of *Arabidopsis thaliana* wild type plants and starchless mutant plants deficient in phosphoglucomutase activity (PGM) were sampled at intervals throughout the day/night cycle. Using principal and independent components analysis, each dataset (metabolites and proteins) displayed discrete characteristics. Compared to the analysis of only metabolites or only proteins, independent components analysis (ICA) of the integrated metabolite/protein dataset resulted in an improved ability to distinguish between WT and PGM plants (first independent component) and, in parallel, to see diurnal variations in both plants (second independent component). Interestingly, levels of photorespiratory intermediates such as glycerate and glycine best characterized phases of diurnal rhythm, and were not influenced by high sugar accumulation in PGM plants. In contrast to WT plants, PGM plants showed an inversely regulated cluster of N-rich amino acid metabolites and carbohydrates, indicating a shift in C/N partitioning. This observation corresponds to altered utilization of urea cycle intermediates in PGM plants suggesting enhanced protein degradation and carbon utilization due to growth inhibition. Among the proteins chloroplastidic GAPDH (At3g26650) was the best discriminator between WT and PGM plants in contrast to the cytosolic isoform (At1g13440) according to the primary effect of mutation located in the chloroplast. The described method is applicable to all kinds of biological systems and enables the unbiased identification of biomarkers embedded in correlative metabolite–protein networks.

KEY WORDS: metabolomics; proteomics; shotgun proteomics; multivariate data analysis; PCA; ICA; unsupervised methods; systems biology; diurnal rhythm.

Abbreviations: AA: ascorbic acid; Ala: alanine; Ara/Xyl: arabinose/xylose; Asn: asparagine; Asp: aspartic acid; BA: benzoic acid; b-Ala: beta-Alanine; CHO(1–12): carbohydrate(1–12); CitA: citric acid; Citn: citrulline; CMA: citramalic acid; Cys: cysteine; EA: ethanolamine; F6P: fructose 6-phosphate; Fru: fructose; Fuc: fucose; FumA: fumaric acid; G1P: glucose 1-phosphate; G6P: glucose 6-phosphate; GA: galactonic acid; GABA: 4-aminobutyric acid; GalOH: galactinol; Glc: glucose; Gln: glutamine; Glu: glutamic acid; Gly: glycine; Glyc: glycerol; GlycA: glyceric acid; HA: hydroxylamine; HyPro: 4-hydroxyproline; IAN: indole-3-acetonitrile; Ile: isoleucine; *iso*-SinA: *iso*-sinapinic acid; Leu: leucine; Lys: lysine; Mal: maltose; MalA: malic acid; Man: mannose; Met: methionine; myo-IN: myo-inositol; Orn/Arg: ornithine/arginine; P: phosphoric acid; PA: propylamine-2,3-diol; pGlu: pyroglutamic acid; Phe: phenylalanine; Pro: proline; Psi: psicose; Put: putrescine; PyrA: pyruvic acid; Raf: raffinose; Rib: ribose; RibA: ribonic acid; SalA: salicylic acid; Ser: serine; SinA: sinapinic acid; Spd: spermidine; Suc: sucrose; SucA: succinic acid; TAmam: tartronic acid 2-(methylaminomethyl); Thr: threonine; ThrA: threonic acid; ThrAL: threonic acid-1,4-lactone; Tre: trehalose; Tyr: tyrosine; UA: uric acid; Ura: uracil; Urea: urea; Val: valine; ADC: arginine decarboxylase; AIH: agmatine iminohydrolase; ARG: arginase; ASL: argininosuccinate lyase; ASS: argininosuccinate synthase; CPA: *N*-carbamoylputrescine amidohydrolase; CPS: carbamoylsynthetase; dSAM: decarboxylated *S*-adenosylmethionine; MTA: 5′-methylthioadenosine; OCT: ornithine carbamoyltransferase; SST: spermidine synthase; PCA: principal components analysis; ICA: independent components analysis.

* To whom correspondence should be addressed.
E-mail: weckwerth@mpimp-golm.mpg.de

1573-3882/05/0400–0109/0 © 2005 Springer Science+Business Media, Inc.

1. Introduction

Metabolite measurements have been used for decades to characterize biological systems (Gerhardt *et al.*, 1987; Nicholson *et al.*, 1999, 2002; Trethewey *et al.*, 1999; Ott *et al.*, 2003; Sumner *et al.*, 2003; Castrillo and Oliver, 2004; Goodacre *et al.*, 2004). With the introduction of novel methods and instrumentation, the comprehensiveness of metabolite detection has reached a new level (Webb *et al.*, 1986; Sauter *et al.*, 1991; Halket *et al.*, 1999; Fiehn, 2000, 2002; Roessner *et al.*, 2000). Recently, we introduced GC-TOF-MS analysis of ultra-complex plant extracts for the detection and quantification of several hundreds of metabolites in a single sample (Weckwerth *et al.*, 2001, 2004a, b). This method provides a hitherto unknown magnitude of metabolite detection due to a unique combination of high resolution gas chromatography with a rapid and sensitive time-of-flight mass analyzer (Watson *et al.*, 1990; Leonard and Sacks, 1999; Veriotti and Sacks, 2001). It further accelerates the analyses thereby increasing sample throughput. These two features make this one of the best methods to describe metabolite dynamics in living systems.

Rapid analysis allows the biological system to be described based on many replicate samples. Independent replicates exhibit high biological variance, even though growth, harvest, and extraction are performed under strictly controlled conditions (Weckwerth, 2003; Weckwerth *et al.*, 2004b). These systems snapshots are a strong and characteristic reflection of the molecular phenotype (Kell and Mendes, 2000; Cooper *et al.*, 2002; Weckwerth *et al.*, 2004b) and can be exploited using multivariate data analysis and supervised machine learning techniques (Cao *et al.*, 1999; Kell, 2002; Goodacre, 2003; Goodacre *et al.*, 2004). For instance, highly-correlated metabolite pairs can be determined and used to visualize "co-regulated" metabolite clusters and metabolic network topologies (Weckwerth *et al.*, 2004a). By comparing differential metabolic networks, such as those of mutant and control plants, regulatory effects are revealed (Weckwerth, 2003; Weckwerth *et al.*, 2004a).

The complexity of regulatory organization inherently implies that molecular phenotypes are not phenomena that can be understood in the context of single gene expressions or perturbations, but rather as the output of interactive biochemical networks (ter Kuile and Westerhoff, 2001) constituted by co-regulated components of transcripts, proteins, and metabolite levels (Weckwerth *et al.*, 2004b). Though not per se containing causality (Wagner, 1997), co-regulation of gene and protein expression analysis and the resulting metabolic phenotype correspond well to our understanding of the causal genotype–phenotype interplay (Cooper *et al.*, 2002; Urbanczyk-Wochniak *et al.*, 2003; Ihmels *et al.*, 2004). From this, it is evident that co-regulation and causal connectivity can be defined best if variables of different levels are analyzed. Gene function is ultimately connected to the corresponding protein action. Proteins carry out all of the catalytic, recognition, structural, and other activities necessary to exert the functional properties associated with the genes. Therefore, proteomics offers the most direct characterization of the function of individual genes at the molecular level.

In the present work, we explore a novel strategy for identifying of time-dependent system regulation and biomarkers using integrative metabolite and protein profiling. The comprehensive profiling of biological samples requires both statistical and novel data-mining tools to reveal significant correlations. Here, an integrative approach is presented founded on independent components analysis (ICA) and revealing an optimized separation of biologically significant processes. Based on the unbiased identification of hundreds of individual compounds, the process enables the identification of characteristic biomarkers in the context of metabolite–protein correlation networks.

2. Materials and methods

2.1. Reagents

Chemicals were purchased from Sigma (Taufkirchen, Germany), except D-sorbitol–^{3} C_6, DL-leucine-2,3,3-d_3, and L-aspartic acid-2,3,3-d_3 which were obtained from Isotech (Miamisburg, USA). Acetonitrile was from J.T. Baker (Deventer, Netherlands), Endoproteinase Lys-C from Boehringer (Mannheim, Germany), and Poroszyme® Immobilized Trypsin from Applied Biosystems (Foster City, USA).

2.2. Plant material and harvest

Arabidopsis thaliana plants Col-0 (wild type) and a plastidic PGM mutant (Caspar *et al.*, 1985) were cultivated simultaneously under identical phytotron conditions set as follows: The temperature and light conditions were 160 μE for 8 h followed by 16 h at 0 μE (darkness). Relative humidity and temperature conditions were set to 70% and 20°C during the light and dark period, respectively.

Plants were harvested at the developmental stage 5.10 (Boyes *et al.*, 2001) at six different time points per day, with 10 different plants per time point and genetic background, respectively. Enzymatic activity was quenched by immediately freezing the plants in liquid nitrogen. Tissues were stored at −80°C until further analysis.

2.3. Extraction procedure and sample preparation for metabolite and protein analysis

Frozen leaf tissue was individually homogenized under liquid nitrogen using a prechilled mortar and pestle. Approximately 20 mg powdered material was used for analysis. Simultaneous extraction of metabolites and proteins from individual plants was performed as described (Weckwerth *et al.*, 2004b) with slight modifications. For metabolite extraction, 1 mL of the extraction mixture containing methanol/chloroform/water (2.5:1:0.5 v:v:v) and 10 μL of an internal standard solution containing 2 mg/mL of each D-sorbitol$-^{13}C_6$, DL-leucine-2,3,3-d_3, and L-aspartic acid-2,3,3-d_3 was added. Soluble metabolites were extracted by mixing the solution at 4°C for 10 min. After centrifugation for 6 min at 20,000 rpm, the supernatant was separated into chloroform and water/methanol phases. The aqueous phase was used for metabolite analysis.

Samples were derivatized by dissolving the dried metabolite pellet in 20 μL of methoxyamine hydrochloride (40 mg/mL pyridine) and shaking the mixture for 90 min at 30°C. After addition of 180 μL of *N*-methyl-*N*-trimethylsilyltrifluoroacetamid (MSTFA), the mixture was incubated at 37°C for 30 min with vigorous shaking. A solution of even-numbered fatty acid methylesters, methylcaprylate (C8-ME), methylcaprate (C10-ME), methyllaurate (C12-ME), methylmyristate (C14-ME), methylpalmitate (C16-ME), methylstearate (C18-ME), methyleicosanoate (C20-ME), methyldocosanoate (C22-ME), lignoceric acid methylester (C24-ME), methylhexacosanoate (C26-ME), methyloctacosanoate (C28-ME), and triacontanoic acid methylester (C30-ME) (each 0.8 mg/mL $CHCl_3$) was spiked into the derivatized sample prior to injection into the GC. Proteins were dissolved from the remaining pellet with 1 mL of freshly prepared protein extraction buffer (8 M urea, 50 mM Tris, 200 mM methylamine, 1% β-mercaptoethanol, pH 7.5). Subsequent protein extraction was performed with water-saturated phenol. Proteins were precipitated with ice-cold acetone at −20°C over night and washed two times with ice-cold methanol to remove residues of β-mercaptoethanol. The dried protein pellets were redissolved in protein extraction buffer and the resulting complex protein mixture was then digested in two steps using endoproteinase Lys-C (1:100) and Poroszyme® Immobilized trypsin according to the manufacturers instructions. The protein digest was desalted with SPEC C18 columns. After lyophilisation the pellet was stored at −20°C until use. For LC-MS/MS shotgun analysis three protein pellets were pooled before enzymatic digestion. All measurements were performed in triplicates.

2.4. GC-TOF-MS analysis

The GC-TOF-MS analysis was performed on an HP 5890 gas chromatograph with deactivated standard spit/splitless liners containing glasswool (Agilent, Böblingen, Germany). One μL sample was injected in the splitless mode at 230°C injector temperature. GC was operated on a MDN-35 capillary, 30 m × 0.32 mm inner diameter, 25 µm film (SUPELCO, Bellefonte, USA), at constant flow of 2 mL/min helium. The temperature program started with 2 min isocratic at 85°C, followed by temperature ramping at 15°C/min to a final temperature of 360°C which was held for 8 min. Data acquisition was performed on a Pegasus II TOF mass spectrometer (LECO, St. Joseph, MI) with an acquisition rate of 20 scans s^{-1} in the mass range of m/z = 85–600.

The obtained data were analyzed at first by defining a reference chromatogram with the maximum number of detected peaks over a signal/noise threshold of 50. Afterwards all chromatograms were matched against the reference with a minimum match factor of 800. Compounds were annotated by retention index and mass spectra comparison to a user defined spectra library. Selected fragment ions specific for each individual metabolite were used for quantification.

2.5. LC-MS shotgun protein analysis

Prior to MS analysis, pellets of protein digests were dissolved in 5% FA. 200 µg per sample are concentrated on a pre-column and subsequently loaded onto a 50 cm silica-based C18 RP monolithic column (Wienkoop *et al.*, 2004a). Elution of the peptides was performed using a 4 h gradient from 100% solvent A (5% acetonitril, 0.1% formic acid in water) to 100% solvent B (90% acetonitril, 0.1% formic acid in water) using the Agilent nano HPLC system (Agilent, Böblingen, Germany) with a flow rate of 400 nL per min. Eluting peptides were analyzed with an LTQ mass spectrometer (Thermo Electron, San Jose) operated in a data-dependent mode. Each full MS scan was followed by three MS/MS scans, in which the three most abundant peptide molecular ions were dynamically selected for collision-induced dissociation (CID) using a normalized collision energy of 35%. The temperature of the heated capillary and electrospray voltage were 150°C and 1.9 kV, respectively. After MS analysis, DTA files were created from raw files and searched against a database (http://www.arabidopsis.org/) using Bioworks 3.1. With DTASelect, a list of identified proteins was obtained using the following criteria: Xcorr: −1 2.0, −2 1.5, −3 3.3 (Peng *et al.*, 2003) for hits with at least two different peptides. For quantitative analysis, CONTRAST was used to compare identified proteins and peptides from different runs (Tabb *et al.*, 2002). These peptides were extracted from the chromatogram and peak areas integrated according to Weckwerth *et al.* (2004b).

2.6. Statistical data analysis

All data were normalized to mg fresh weight and stable isotope-labeled standard compounds. Statistical tests were performed in MATLAB 6.5 (Mathworks, Natick, MA) on the basis of log-transformed data.

The threshold value chosen for the Pearson correlation was 0.85. Significance levels for Pearson correlations r were computed depending on the number of metabolite pairs n (equal to the number of samples) by calculating t-scores given by $t = r(n - 2)^{0,5} /(1 - r^2)^{0,5}$ and controlled for potential impact of outliers by robust fit assessments. Critical t-score was set corresponding to the common used p-value of 0.001. As additional criterion for correlations accepted we defined a variance threshold based on the results of the measurements of replicates. Correlations were found relevant, if the coefficient of variation of metabolites involved was above 30% and thus exceeded the analytical error (see figure 1). Thus correlations between pairs of metabolites were considered only if these requirements corresponded.

Independent components analysis (ICA) was applied in combination with principal components analysis (PCA) as pre-processing and the measure of kurtosis as evaluation criterion. The dimensionality of the data was first reduced by PCA to a set of three principal components (PCs). ICA was then applied to this reduced dataset and the extracted independent components were ranked by the kurtosis measure. The contributions of each metabolite/protein to a independent component can be obtained by combining the transformation matrix W of PCA with the transformation matrix V of ICA to a direct transformation $U = W{*}V$. The elements of the ith vector in U represent the individual contributions to the ith independent component (ICi).

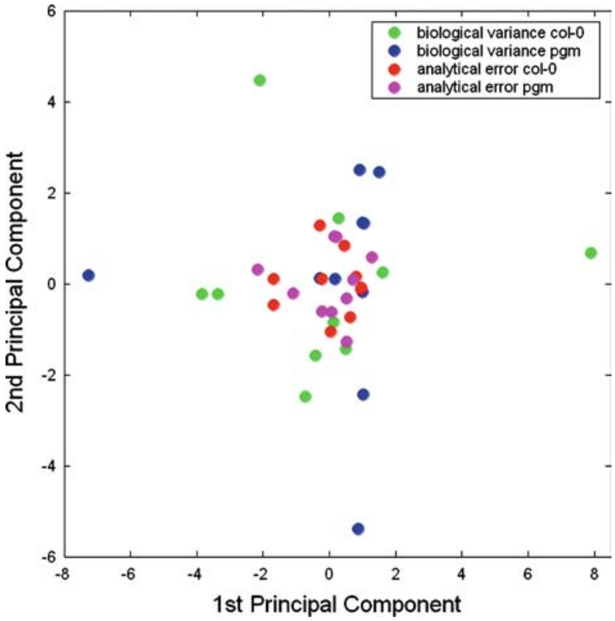

Figure 1. PCA of log-transformed metabolite data from samples harvested after 4 h of light for *Arabidopsis thaliana* wild type and mutant plants. Each measurement (containing in total 80 quantified metabolites) is represented by a dot. The analytical error including the technical error occurring during extraction, derivatization, and measurement is diminutive compared to measurements of individual plant samples representing the biological variance. The biological variability is in multiple excess of the variability caused by technical limitations.

For more details see Scholz *et al.* (2004). The ICA algorithm used in this article is CuBICA4 (Blaschke and Wiskott, 2004).

3. Results and discussion

3.1. Combined metabolite and protein extraction

When performing simultaneous analyses of intracellular metabolites and proteins extracted from a biological sample, several aspects of sample preparation are crucial. First, all enzymatic activity must be quenched as rapidly as possible to avoid protein degradation and to stop metabolism. Second, a full spectrum of metabolites must be extracted to get a clear and full picture of the metabolic state of the organism. Third, the extraction procedure must be rapid and efficient to avoid undesired post-extraction reactions of the metabolites (such as oxidation) that could interfere with unambiguous metabolite identification and quantification. This holds especially true when profiling plant tissues, where the harvesting process and sample preparation must be tightly controlled to ensure high accuracy (Weckwerth, 2003). Recently, we developed an integrative extraction protocol for the co-extraction of metabolites and proteins from the same biological sample using an ice-cold methanol/chloroform/water mixture for rapid quenching and for metabolite extraction (Weckwerth *et al.*, 2004b).

In this procedure, urea was used as a chaotropic reagent to denature and solubilize proteins during extraction and enzymatic digestion prior to LC-MS analysis. However, urea is susceptible to decomposition, yielding ammonium cyanate, which leads to artificial modifications of cysteine, lysine, or arginine side chains. This may result in false identifications of proteins via peptide sequencing. To overcome this limitation we modified the method slightly by using a protein extraction and digestion buffer that contains methylamine, a scavenger of cyanate (Lippincott and Apostol, 1999).

3.2. GC-TOF-MS for metabolite profiling in the starchless *A. thaliana* mutant PGM

The well-studied starchless *A. thaliana* mutant PGM was chosen as a model system. PGM plants lack plastidic phosphoglucomutase, an enzyme that catalyzes the reversible interconversion of glucose 1-phosphate and glucose 6-phosphate. One of the distinguishing characteristics of the mutant is that it accumulates relatively large quantities of sucrose, glucose, and fructose in both leaf and stem tissue (Caspar *et al.*, 1985). Thus, pleiotropic effects at the metabolite level are expected to be pronounced. Sugar accumulation in PGM plants leads to an increase in the kinase activity associated with sucrose-phosphate synthase (SPS), a highly regulated enzyme that catalyzes the penultimate step in sucrose synthesis in plants. This increase may correspond to a downregulation of SPS within the pathway (Winter and Huber, 2000; Glinski *et al.*, 2003). Studies at the transcriptional level revealed widespread changes in the PGM plants compared to the WT (Gibon *et al.*, 2004a, b; Thimm *et al.*, 2004).

The plants were grown in an 8 h light/16 h dark regime. Samples for metabolic profiling were prepared as described in the experimental section. For PGM samples that were harvested in the middle of the light period, significantly increased amounts of carbohydrates such as glucose, sucrose, and fructose were observed that interfered with the dynamic range of the detector. Therefore, measurements of dilution series were performed. The analysis of the chromatographic runs was performed using the Pegasus software package (Leco). Spectra were compared to a user-defined database, containing the specific fragmentation patterns of about 600 metabolites (Weckwerth *et al.*, 2004a). To provide additional criteria for identification and to overcome variations in column temperature and flow-rate, the retention index procedure described by Kovats (Kovats, 1958) was performed with a few modifications. Due to the occurrence of *n*-alkanes in lipid metabolism of higher plants (Collister *et al.*, 1994), even-numbered fatty acid methyl esters (FAMEs) ranging from C_8 to C_{30} were used as retention index markers. Specific retention indices calculated with the help of FAMEs were also included in the user library. Therefore, two criteria were used for unambiguous compound identification: (i) the metabolite-specific fragmentation pattern and (ii) the relative retention time corrected by Kovats indices.

To get statistically relevant data it is necessary to perform replicate measurements. From repeat measurements of the same sample we determined the "analytical error" as the sum of all possible technical errors that occur during the extraction, derivatization, GC-TOF-MS analysis, and data evaluation of each individual metabolite. For this purpose pooled plant leaf materials from *A. thaliana* Col-0 WT and PGM mutants harvested in the middle of the light and dark periods, respectively, were extracted, derivatized, and analyzed in 10 replicates.

For accurate quantification the use of internal standards is required. However, in metabolite profiling the adoption of stable isotope-labeled standards for each compound to be identified is critical because of an undesired increase in sample complexity. Therefore, the applicability of three different stable isotope-labeled internal standards for the quantification of polar metabolites was settled on. For that, all metabolites identified were normalized to these stable isotope-labeled standards. As a result, sugars and sugar acids could be best quantified via D-sorbitol–$^{13}C_6$, amino acids via DL-leucine-2,3,3-d_3, and TCA cycle intermediates via L-aspartic acid-2,3,3-d_3. The coefficient of variation (CV) for the quantification of sugars and sugar acids was below 20% RSD; for amino acids and TCA cycle intermediates a maximum CV of 24% was determined. Moreover, the analytical errors determined for the analysis of samples harvested during the dark did not differ significantly from the analytical errors obtained for samples harvested during the light period (data not shown). A PCA analysis of the results is shown in figure 1. The biological fluctuation of independent samples exceeds the analytical precision several fold (Weckwerth, 2003; Weckwerth *et al.*, 2004b). Based on this approach we defined the variance threshold for significant metabolite correlations. The variances and the correlations or their differences are the basis of multivariate data analysis (see below).

Samples were taken throughout the 8 h/16 h day/night rhythm. As expected, in the PGM plants these measurements revealed strongly increased sugar contents during light periods (see figure 2a–c) (Caspar *et al.*, 1985). From these data, metabolite correlation networks were constructed (Weckwerth and Fiehn, 2002; Weckwerth, 2003; Weckwerth *et al.*, 2004a). Significant differences were immediately visible between WT and PGM plants (see figure 3). For example, the increased sugar levels tended to show a strong correlation in the PGM plants whereas no correlation was apparent in the WT. It is known that the PGM plants have a strongly increased acid invertase activity converting sucrose directly to glucose and fructose (Caspar *et al.*, 1985). Thus, this is a good example to explain the network topology of the correlations on a biochemical level. However, although differential network analysis reveals differences in biochemical regulation as discussed recently (Steuer *et al.*, 2003; Weckwerth, 2003; Weckwerth *et al.*, 2004a) correlations per se give only hints to altered pathway activities and cannot be explained straightforward (Steuer *et al.*, 2003; Camacho *et al.*, 2005). Another, less expected finding was a highly correlated amino acid and urea cycle intermediate cluster (see also below) in PGM plants not found in the WT (see figures 3 and 6). This cluster points to altered C/N partitioning, which has been hinted at previously in studies of transcriptional regulation (Gibon *et al.*, 2004b; Thimm *et al.*, 2004). The involvement of the urea cycle in this process could point to enhanced protein degradation and carbon utilization. However, this is only conjecture.

The correlation matrix combining all time points immediately reveals significant differences between PGM and WT plants (see figure 4). An inversely co-regulated cluster of amino acids is connected to most of the sugar compounds. Again, this is a clear indication that sugar accumulation in PGM plants is regulated inversely to nitrogen metabolism. Intriguingly, this is in close agreement with studies on potato tubers (Roessner-Tunali *et al.*, 2003).

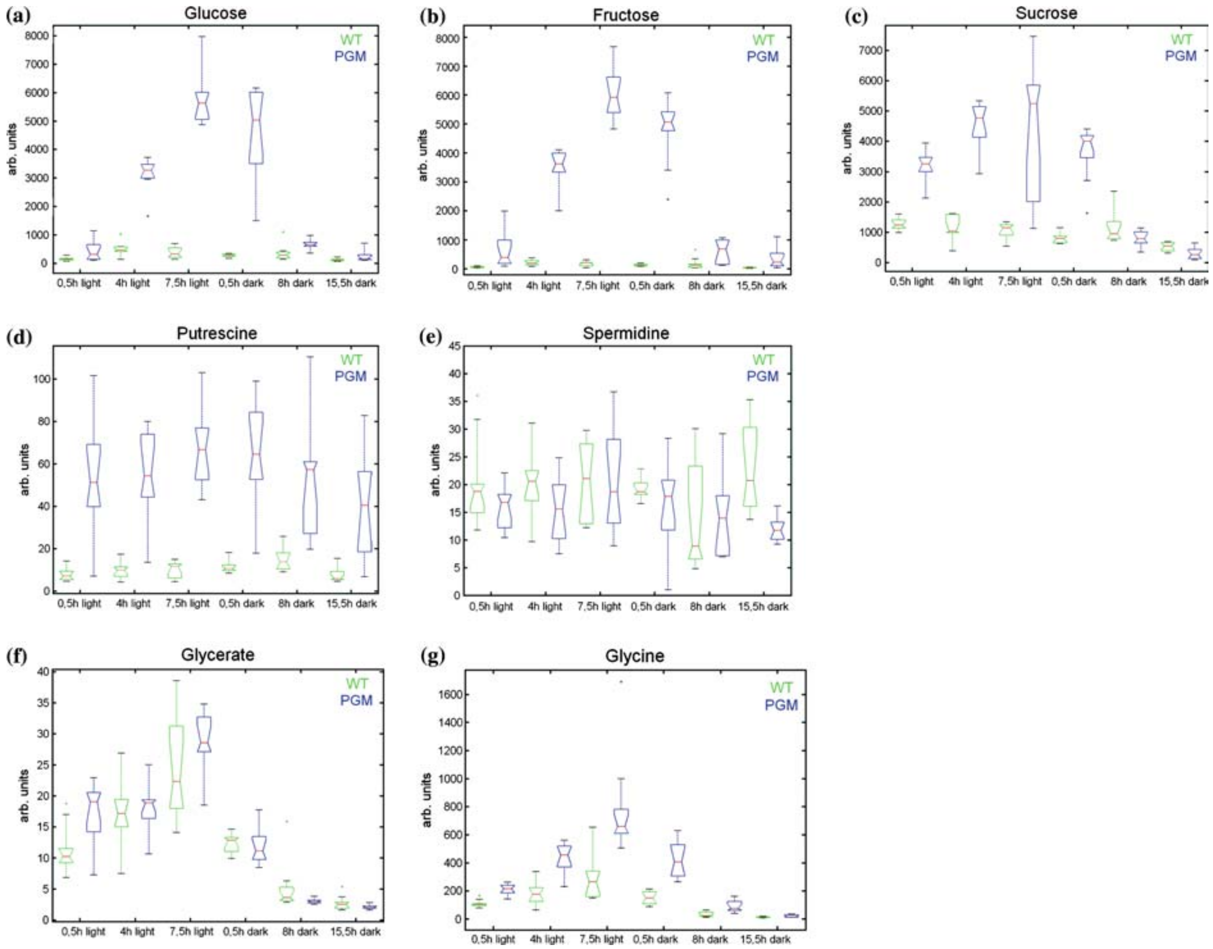

Figure 2. Box plots of selected metabolites representative of samples collected from WT and PGM plants at different time points (8 h day/16 h night) (for details see text). (a) Glucose, (b) Fructose, (c) Sucrose, (d) Putrescine (e) Spermidine, (f) Glycerate, (g) Glycine.

3.3. LC-MS shotgun protein analysis as a tool for rapid identification and quantification of proteins

High throughput GC/MS metabolite profiling methods are a prerequisite for systems biology, enabling the analysis of in vivo dynamics in real-world samples (Weckwerth, 2003). In this study, we explored the possibility of achieving similar throughput with LC-MS-based protein profiling (Wienkoop *et al*., 2004b). Highly complex tryptic peptide mixtures of non-fractionated protein samples were analyzed using one-dimensional reversed phase chromatography coupled to mass spectrometry. Long monolithic columns providing high peak resolution capacity were employed (Wienkoop *et al*., 2004a). Due to this restriction the scale of identified and quantified proteins was relatively small (40 proteins; see table 1). However, the question arose whether the data

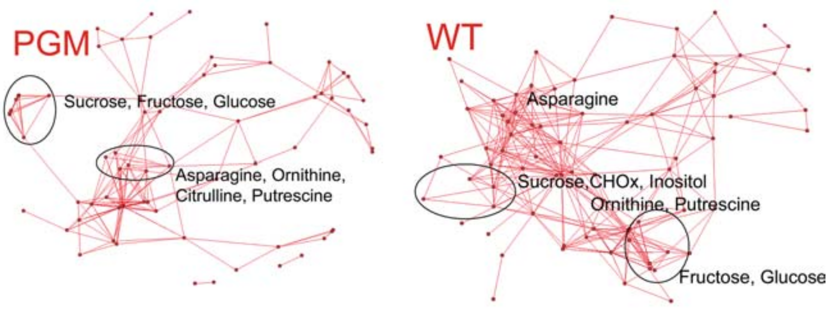

Figure 3. Metabolite correlation networks in PGM and WT plants at the 4 h day time point. For a detailed description of criteria used to define correlations as significant see section 2.6. Differences in the network structure were pronounced (for details see text and Weckwerth, 2003; Weckwerth *et al*., 2004a).

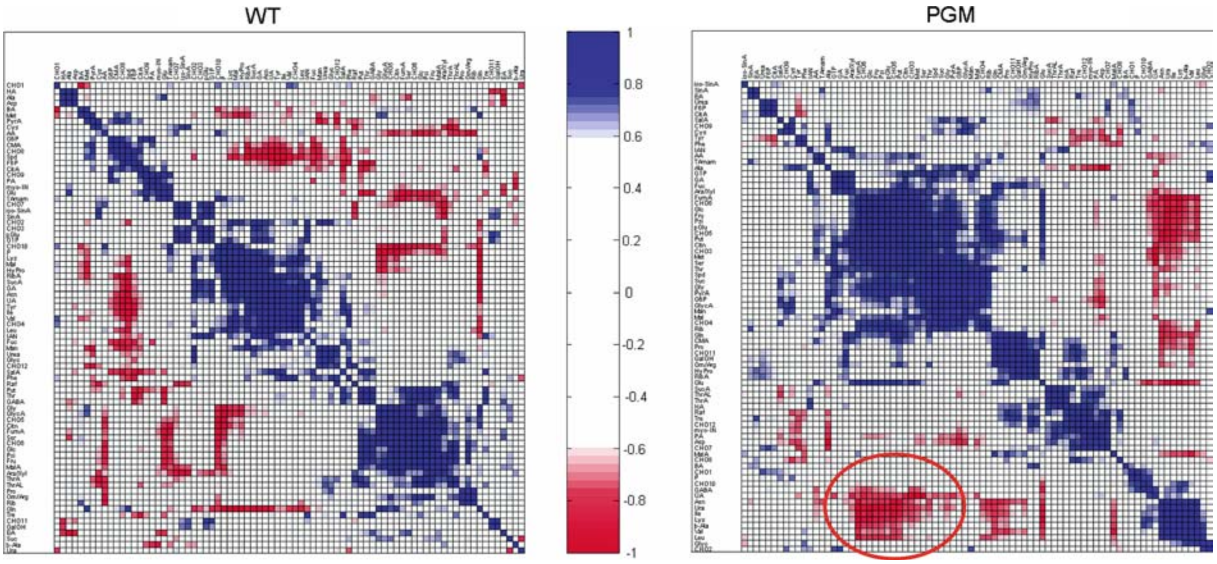

Figure 4. Correlation matrix averaging all data points throughout a diurnal rhythm. The inversely correlated cluster of N-rich amino acids and sugars for PGM plants in contrast to the WT plants (circled) is clearly visible. This effect was pronounced in the PGM plants where it was most likely triggered by sugar accumulation and carbon mobilization (see text for detailed discussion).

reveal interesting protein marker that make it possible to distinguish between PGM and WT plants as well as between day and night samples. Indeed, figure 5a shows a protein PCA plot that nicely distinguishes between mutant and WT plants. Changes due to diurnal rhythm, however, are only visible for the mutant. Taken together, the protein data mirror the biology behind the experimental setup.

3.4. *Integrative metabolite/protein multivariate data analysis improves phenotyping and enables biomarker selection embedded in dynamic metabolite–protein networks*

Supervised analysis methods such as discriminatory analysis tend to enhance non-specific effects in ''noisy'' data with small sample numbers and many variables. Therefore, we have chosen PCA and ICA, called unsupervised techniques, to identify inherent biological characteristics independently of experimental background information. PCA is a well-established technique for dimensionality reduction and visualization (Diamantaras and Kung, 1996). In the field of metabolomics, PCA has become a popular tool for visualizing datasets and for extracting relevant information (Fiehn *et al*., 2000; Viant, 2003; Goodacre *et al*., 2004). However, PCA is only powerful if the biological question is related to the highest variance in the dataset. Different techniques exist to extend PCA. Several extensions are non-linear PCA (Scholz and Vigário, 2002), or locally linear embedding (Roweis and Saul, 2000). Due to the limited number of samples in high-dimensional datasets, linear alternatives might be more reliable. A promising linear technique is ICA. Similar to PCA, ICA provides a set of components which can be seen as new variables. The components are linear combinations of the original variables (metabolites/proteins) and represent specific directions in the original data space. In contrast to PCA these components are constructed in order to minimize the dependence and are therefore termed independent components (ICs). Independence is a stronger condition than non-correlation in PCA and gives often more meaningful components. The components of ICA are not required to be orthogonal. There is a large variety of methods for performing ICA (Hyvärinen *et al*., 2001; Cichocki and Amari, 2002). Usually ICA is applied to datasets having a large number of samples and only a small number of variables. Molecular data, in contrast, impose a large number of variables (metabolites, proteins) compared to a relatively small number of samples. Applying ICA directly to this high-dimensional dataset is questionable and the results are usually of no practical relevance. Furthermore, ICA extracts as many components as the dataset has dimensions. The components have by default no order, and hence a criterion is needed to rank the components. For that we use the kurtosis measure which is a classical measure of non-Gaussianity. Consequently, to apply ICA, we first reduce the dimensionality by PCA, thereby maintaining all of the relevant variances. Then ICA is applied to this reduced dataset and the extracted independent components are ranked by the kurtosis measure. For more details see Scholz *et al*. (2004). We applied both PCA and ICA to the metabolite, protein and an integrated dataset containing all the metabolite and protein data together. Data normalization enabled the comparison of profiling data originating from different instruments such as GC-TOF-MS data and iontrap data (LCQ). The quantitative data received here do not represent absolute concentrations,

Table 1
Proteins identified and quantified via peak integration

AGI-Code	Annotation
At1g06680.1	68414.m00708 photosystem II oxygen-evolving complex 23 (OEC23) JBC 14:211–238 (2002); identical to 23 kDa polypeptide of oxygen-evolving comlex (OEC) GB:CAA66785 GI:1769905 [*Arabidopsis thaliana*]
At1g13440.1	68414.m01570 glyceraldehyde 3-phosphate dehydrogenase, cytosolic, putative/NAD-dependent glyceraldehyde-3-phosphate dehydrogenase, putative very strong similarity to SP_P25858 Glyceraldehyde 3-phosphate dehydrogenase, cytosolic (EC 1.2.1.12) *Arabidopsis thaliana*; contains Pfam profiles PF02800: Glyceraldehyde 3-phosphate dehydrogenase C-terminal domain, PF00044: Glyceraldehyde 3-phosphate dehydrogenase NAD binding domain
At1g13930.1	68414.m01635 expressed protein weakly similar to drought-induced protein SDi-6 (PIR:S71562) common sunflower (fragment)
At1g19100.1	68414.m02376 ATP-binding region, ATPase-like domain-containing protein-related low similarity to microrchidia [Homo sapiens] GI:5410257; contains non-consensus splice site (GC) at intron 8
At1g51965.1	68414.m05859 pentatricopeptide (PPR) repeat-containing protein contains Pfam profile PF01535: PPR repeat
At1g67090.1	68414.m07629 ribulose bisphosphate carboxylase small chain 1A/RuBisCO small subunit 1A (RBCS-1A) (ATS1A) identical to SP_P10795 Ribulose bisphosphate carboxylase small chain 1A, chloroplast precursor (EC 4.1.1.39) (RuBisCO small subunit 1A) *Arabidopsis thaliana*
At2g21170.1	68415.m02511 triosephosphate isomerase, chloroplast, putative similar to Triosephosphate isomerase, chloroplast precursor: SP|P48496 from *Spinacia oleracea*, SP|P46225 from *Secale cereale*
At2g28190.1	68415.m03423 superoxide dismutase [Cu–Zn], chloroplast (SODCP)/copper/zinc superoxide dismutase (CSD2) identical to GP:3273753:AF061519
At2g30790.1	68415.m03754 photosystem II oxygen-evolving complex 23, putative expression not detected; similar to SP_O49344 (GI:28800560 (OEC23) *Arabidopsis*; Non-identical EST and protein matches suggested a possible frameshift in exon 1 (a 4 base deletion between 73745 and 73746) and a different start for exon 2 (base 73645).
At2g35370.1	68415.m04336 glycine cleavage system H protein 1, mitochondrial (GDCSH) (GCDH) identical to SP|P25855 Glycine cleavage system H protein 1, mitochondrial precursor *Arabidopsis thaliana*
At2g37220.1	68415.m04566 29 kDa ribonucleoprotein, chloroplast, putative/RNA-binding protein cp29, putative similar to SP|Q43349 29 kDa ribonucleoprotein, chloroplast precursor (RNA-binding protein cp29) *Arabidopsis thaliana*
At2g39730.1	68415.m04877 ribulose bisphosphate carboxylase/oxygenase activase/RuBisCO activase identical to SWISS-PROT:P10896 ribulose bisphosphate carboxylase/oxygenase activase, chloroplast precursor (RuBisCO activase, RA)[*Arabidopsis thaliana*]
At3g01500.1	68416.m00074 carbonic anhydrase 1, chloroplast/carbonate dehydratase 1 (CA1) nearly identical to SP|P27140 Carbonic anhydrase, chloroplast precursor (EC 4.2.1.1) (Carbonate dehydratase) *Arabidopsis thaliana*
At3g14210.1	68416.m01796 myrosinase-associated protein, putative similar to GB:CAA71238 from [*Brassica napus*]; contains Pfam profile:PF00657 Lipase/Acylhydrolase with GDSL-like motif
At3g15360.1	68416.m01948 thioredoxin M-type 4, chloroplast (TRX-M4) nearly identical to SP_Q9SEU6 Thioredoxin M-type 4, chloroplast precursor (TRX-M4) *Arabidopsis thaliana*
At3g16890.1	68416.m02159 pentatricopeptide (PPR) repeat-containing protein contains Pfam profile PF01535: PPR repeat
At3g26650.1	68416.m03330 glyceraldehyde 3-phosphate dehydrogenase A, chloroplast (GAPA)/NADP-dependent glyceraldehydephosphate dehydrogenase subunit A identical to SP_P25856 Glyceraldehyde 3-phosphate dehydrogenase A, chloroplast precursor (EC 1.2.1.13) (NADP-dependent glyceraldehydephosphate dehydrogenase subunit A) *Arabidopsis thaliana*
At3g27690.1	68416.m03457 chlorophyll A–B binding protein (LHCB2:4) nearly identical to Lhcb2 protein [*Arabidopsis thaliana*] GI:4741950; similar to chlorophyll A–B binding protein 151 precursor (LHCP) GB:P27518 from [*Gossypium hirsutum*]; contains Pfam PF00504: Chlorophyll A–B binding protein
At3g27830.1	68416.m03471 50S ribosomal protein L12–1, chloroplast (CL12-A) identical to ribosomal protein L12 GB:X68046 [*Arabidopsis thaliana*] (J. Biol. Chem. 269 (10), 7330–7336 (1994))
At3g47070.1	68416.m05111 expressed protein
At3g50820.1	68416.m05565 oxygen-evolving enhancer protein, chloroplast, putative/33 kDa subunit of oxygen evolving system of photosystem II, putative (PSBO2) identical to SP:Q9S841 Oxygen-evolving enhancer protein 1–2, chloroplast precursor (OEE1) [*Arabidopsis thaliana*]; strong similarity to SP_P23321 Oxygen-evolving enhancer protein 1–1, chloroplast precursor (OEE1) (33 kDa subunit of oxygen evolving system of photosystem II) (OEC 33 kDa subunit) (33 kDa thylakoid membrane protein) *Arabidopsis thaliana*
At3g55800.1	68416.m06200 sedoheptulose-1,7-bisphosphatase, chloroplast/sedoheptulose-bisphosphatase identical to SP|P46283 Sedoheptulose-1,7-bisphosphatase, chloroplast precursor (EC 3.1.3.37) (Sedoheptulose-bisphosphatase) (SBPASE) (SED(1,7)P2ASE) *Arabidopsis thaliana*
At4g02530.1	68417.m00346 chloroplast thylakoid lumen protein SP:022773; TL16_ARATH
At4g03280.1	68417.m00447 cytochrome B6-F complex iron–sulfur subunit, chloroplast/Rieske iron–sulfur protein/plastoquinol-plastocyanin reductase (petC) identical to gi:9843639; identical to cDNA rieske iron–sulfur protein precursor (petC) GI:5725449
At4g05180.1	68417.m00778 oxygen-evolving enhancer protein 3, chloroplast, putative (PSBQ2) identical to SP_Q41932 Oxygen-evolving enhancer protein 3–2, chloroplast precursor (OEE3) (16 kDa subunit of oxygen evolving system of photosystem II) (OEC 16 kDa subunit) *Arabidopsis thaliana*; similar to SP_P12301 Oxygen-evolving enhancer protein 3, chloroplast precursor (OEE3) (16 kDa subunit of oxygen evolving system of photosystem II) (OEC 16 kDa subunit) Spinacia oleracea; contains Pfam profile PF05757: Oxygen evolving enhancer protein 3 (PsbQ)
At4g05320.3	68417.m00812 polyubiquitin (UBQ10) (SEN3) senescence-associated protein; identical to GI:870791
At4g10340.1	68417.m01699 chlorophyll A–B binding protein CP26, chloroplast/light-harvesting complex II protein 5/LHCIIc (LHCB5) identical to SP_Q9XF89 Chlorophyll A/B-binding protein CP26, chloroplast precursor (Light-harvesting complex II protein 5) (LHCB5) (LHCIIc) *Arabidopsis thaliana*; contains Pfam profile: PF00504 chlorophyll A–B binding protein; chlorophyll a/b-binding protein CP26 in PS II, Brassica juncea, gb:X95727

Table 1
Continued

AGI-Code	Annotation
At4g21280.1	68417.m03075 oxygen-evolving enhancer protein 3, chloroplast, putative (PSBQ1) (PSBQ) identical to SP_Q9XFT3 Oxygen-evolving enhancer protein 3–1, chloroplast precursor (OEE3) (16 kDa subunit of oxygen evolving system of photosystem II) (OEC 16 kDa subunit) *Arabidopsis thaliana*; similar to SP_P12301 Oxygen-evolving enhancer protein 3, chloroplast precursor (OEE3) (16 kDa subunit of oxygen evolving system of photosystem II) (OEC 16 kDa subunit) Spinacia oleracea; contains Pfam profile PF05757: Oxygen evolving enhancer protein 3 (PsbQ)
At4g23920.1	68417.m03440 UDP-glucose 4-epimerase, putative/UDP-galactose 4-epimerase, putative/Galactowaldenase, putative similar to UDP-galactose 4-epimerase from Arabidopsis thaliana SP_Q42605, Cyamopsis tetragonoloba GI:3021357 [AJ005082]
At4g28660.1	68417.m04096 photosystem II reaction center W (PsbW) family protein contains Pfam profile: PF03912 photosystem II reaction center W protein, PsbW
At4g28750.1	68417.m04111 photosystem I reaction center subunit IV, chloroplast, putative/PSI-E, putative (PSAE1) identical to SP|Q9S831; similar to SP|P12354 Photosystem I reaction center subunit IV, chloroplast precursor (PSI-E) Spinacia oleracea; contains Pfam profile PF02427: Photosystem I reaction center subunit IV/PsaE
At5g06240.1	68418.m00697 expressed protein
At5g18740.1	68418.m02224 expressed protein predicted proteins – *Arabidopsis thaliana*; expression supported by MPSS
At5g24430.1	68418.m02879 calcium-dependent protein kinase, putative/CDPK, putative similar to calcium/calmodulin-dependent protein kinase CaMK1 [*Nicotiana tabacum*] gi_16904222_gb_AAL30818
At5g25980.1	68418.m03090 glycosyl hydrolase family 1 protein contains Pfam PF00232: Glycosyl hydrolase family 1 domain; TIGRFAM TIGR01233: 6-phospho-beta-galactosidase; identical to thioglucosidase (GI:871992) [*Arabidopsis thaliana*]; similar to myrosinase precursor (EC 3.2.3.1)(Sinigrinase) (Thioglucosidase) SP_P37702 from [*Arabidopsis thaliana*]
At5g47020.1	68418.m05795 glycine-rich protein strong similarity to unknown protein (emb_CAB87688.1)
At5g66570.1	68418.m08392 oxygen-evolving enhancer protein 1–1, chloroplast/33 kDa subunit of oxygen evolving system of photosystem II (PSBO1) (PSBO) identical to SP:P23321 Oxygen-evolving enhancer protein 1–1, chloroplast precursor (OEE1) (33 kDa subunit of oxygen evolving system of photosystem II) (OEC 33 kDa subunit) (33 kDa thylakoid membrane protein) [*Arabidopsis thaliana*]
ATPB_ARATH	(P19366) ATP synthase beta chain (EC 3.6.1.34)
PSAC_ARATH	(P25252) Photosystem I iron–sulfur center (Photosystem I subunit VII) (9 kDa polypeptide) (PSI-C)
RBL_ARATH	(O03042) Ribulose bisphosphate carboxylase large chain precursor (EC 4.1.1.39) (RuBisCO large subunit)

but arbitrary units. To ensure a comparable scale the measurements of each metabolite and protein were first divided by the median over all measurements of one metabolite or protein. This allows a direct comparison of different metabolite or protein levels. A high relative change in concentration is then still represented by a high variance. Subsequent log-transformation increases the importance of low valued metabolites and compresses the upper end of measurement scale and thus leads to more balanced variables. Furthermore, the

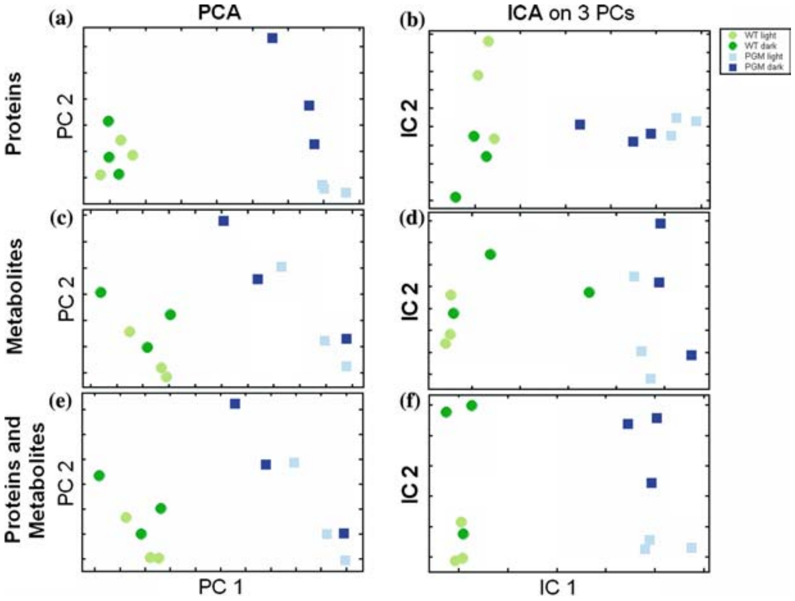

Figure 5. PCA and ICA of the protein data, metabolite data, and the integrative protein/metabolite data matrix. The clearest unsupervised discrimination among the four conditions (PGM, WT, day, night) was found with ICA on the integrative dataset. One data point belonging to WT night was, however, not separated. (a) PCA proteins, (b) ICA proteins, (c) PCA metabolites, (d) ICA metabolites, (e) PCA metabolites/proteins, (f) ICA metabolites/proteins.

influence of potentially outliers is reduced. The often used standardization to 0 mean and 1 variance treats each variable equally; as a result this transformation ignores the experimental variation. Therefore, we decided not to use this so-called *z*-transformation. Application of PCA/ICA to the protein data demonstrated that the projections were able to distinguish among the four experimental conditions (WT, PGM, day, and night; figure 5a, b). This is a valuable demonstration that the relatively small dataset of 40 quantified proteins is able to unravel the experimental and biological design. The diurnal rhythm in the WT, however, is observed in neither the ICA nor the PCA plot of the proteins.

The metabolites show a different grouping compared to the protein data. PGM and WT are clearly distinguishable, but the diurnal rhythm exhibits a quite unexpected behavior. Some metabolites exhibit a time delay into the next period of light or darkness before being metabolized to the representative homeostatic state. Thus, the three day and night points have a strong overlap but show the expected order according to the sampling time (see figure 5c, d). However, the separation is not as clear as that observed for the protein data.

Discrimination amongst all four conditions (day, night, PGM, WT) is best for the combined analysis of metabolites and proteins (see figure 5e, f). ICA has the highest discriminating power: the separation of PGM and WT plants as well as the day/night separation of the diurnal rhythm is improved (see figure 5f). Most important, the different biological states are assigned unambiguously to different independent components: the separation of PGM/WT on the first independent component and the diurnal rhythm on the second independent component. For each component the metabolites and proteins with the highest contribution to the component can be identified, as shown in table 2.

In ICA the metabolite variances are more balanced with protein variances (data not shown) yielding an opportunity to define proteins as biomarkers in a context of dynamic metabolite networks. This is an important extension of existing methods where protein differences are ranked based on their quantitative changes to a reference sample. Two classes of proteins displayed high factors: First, components of the photosynthetic machinery (PSAC_ARATH; At3g27690.1; At4g03280.1; At2g30790.1) and second, proteins of the Calvin cycle (At3g26650.1; RBL_ARATH) (see table 2). However, several proteins having a strong dominance in the loadings do not belong to either class, for instance an expressed protein (At1g13930.1) with weak homology to drought-induced protein in sunflower and unknown function. Since this protein contributes more to the separation of PGM/WT than to the diurnal rhythm it is a primary candidate for further functional studies.

Table 2

Factor loadings of the ICA derived from the reduced data set obtained by PCA

ICA: loadings IC 1 (PGM/WT separation)		ICA: loadings IC 2 (diurnal cycle)	
0.07	Glyc	−0.17	Gly
0.04	Asn	−0.16	GlycA
−0.04	PSAC_ARATH	−0.1	Rib
−0.04	At3g26650.1	0.09	Man
0.04	Orn/Arg	0.08	Pro
0.03	Put	0.07	Urea
0.03	At3g27690.1	0.07	SucA
0.03	Gly	−0.07	At5g25980.1
−0.03	At1g13930.1	−0.07	At3g27690.1
0.02	Fru	−0.06	At2g30790.1
0.02	Lys	−0.06	Suc
−0.02	FumA	−0.06	PyrA
0.02	Ile	0.05	Glc
0.02	SalA	−0.05	At1g13930.1
0.02	IAN	−0.05	At3g27830.1
0.02	Citn	−0.05	Met
−0.02	At4g03280.1	0.05	PSAC_ARATH
−0.02	Pro	0.04	Glu
−0.02	At4g05320.3	−0.04	At4g10340.1
0.02	b-Ala	0.04	IAN
−0.02	At4g02530.1	−0.04	At1g06680.1
0.02	Phe	0.04	At3g16890.1
0.02	Psi	0.03	RBL_ARATH
−0.02	At2g30790.1	0.03	myo-IN
0.02	Val	0.03	Asn
−0.02	Gln	−0.03	G6P
0.02	CHO10	0.03	Asp
−0.02	At1g13440.1	−0.03	ATPB_ARATH
0.01	At4g28750.1	−0.03	Ala
−0.01	Glc	0.02	AA
0.01	CHO9	0.02	At4g03280.1
−0.01	CHO8	0.02	At3g26650.1
−0.01	MalA	0.02	pGlu
−0.01	At1g06680.1	−0.02	Ser
0.01	At3g55800.1	0.02	Glyc
0.01	CHO7	0.02	At1g67090.1
0.01	Ala	−0.02	Citn
0.01	Met	0.02	CHO7
0.01	At5g25980.1	0.02	At1g13440.1
−0.01	SinA	0.02	Asc
−0.01	Man	0.02	ThrAL
0.01	CHO1	0.02	P
0.01	At4g10340.1	−0.02	Spd
−0.01	myo-IN	0.02	GA
−0.01	At3g16890.1	...	
0.01	Urea		

Several metabolites displayed high weights on the components. As expected, sugar accumulation (sucrose, fructose and glucose – the causal effect of PGM deficiency) contributed strongly to the separation of PGM and WT as well as to the grouping of the day/night rhythm. However, the strongest weights for the diurnal rhythm are found for the photorespiratory intermediates glycerate and glycine (Geiger and Servaites, 1994) (see table 2 and figure 2f, g), indicating that photorespiration in addition to CO_2 fixation has a strong impact in the day/night rhythm and is not strongly altered by sugar accumulation or carbon utilization in the PGM plants (Gibon *et al*., 2004b).

There are many other metabolites that unexpectedly show a very different behavior in the WT compared to PGM, for instance urea cycle intermediates, polyamines (see also figure 2d, e) and other amino acids such as proline and glutamate with high loadings (see table 2). These differences directly point to altered pathways in PGM plants (see figure 6) with respect to the control plant as discussed above.

Figure 7 is a demonstration of the pronounced effect of the diurnal rhythm found in the PGM plants based on the integrative metabolite/protein dataset. The diurnal trajectory in the PCA plot which denotes for the behavior of the plant during a day/night cycle is seen easily.

Summarizing, multivariate data analysis of the integrative metabolite–protein data matrix enables the visualization of inherent time-dependent biological characteristics and, consequently, the identification of the most discriminatory metabolites and proteins embedded in a dynamic network of correlations.

4. Concluding remarks

Here we propose a process that integrates quantitative metabolite and protein data at a systems level. Application of multivariate data analysis demonstrated that extraction of inherent biological information is improved using such integrative data. Therefore, we think that the method is useful for future diagnostic technology and for the identification of biomarkers embedded in dynamic biochemical networks. The exploitation of biological variance and amplitudes of fluctuation of independent samples – in contrast to averaging replicates – will be valuable for characterizing biological systems. Our efforts are now directed towards extending the protein profiling method, so as to have representative protein candidates for all kinds of metabolic pathways in hand.

Acknowledgments

We thank Megan McKenzie for revising the manuscript. We thank the Max Planck Society for financial support and Mark Stitt, Mirko Glinski, Oliver Fiehn,

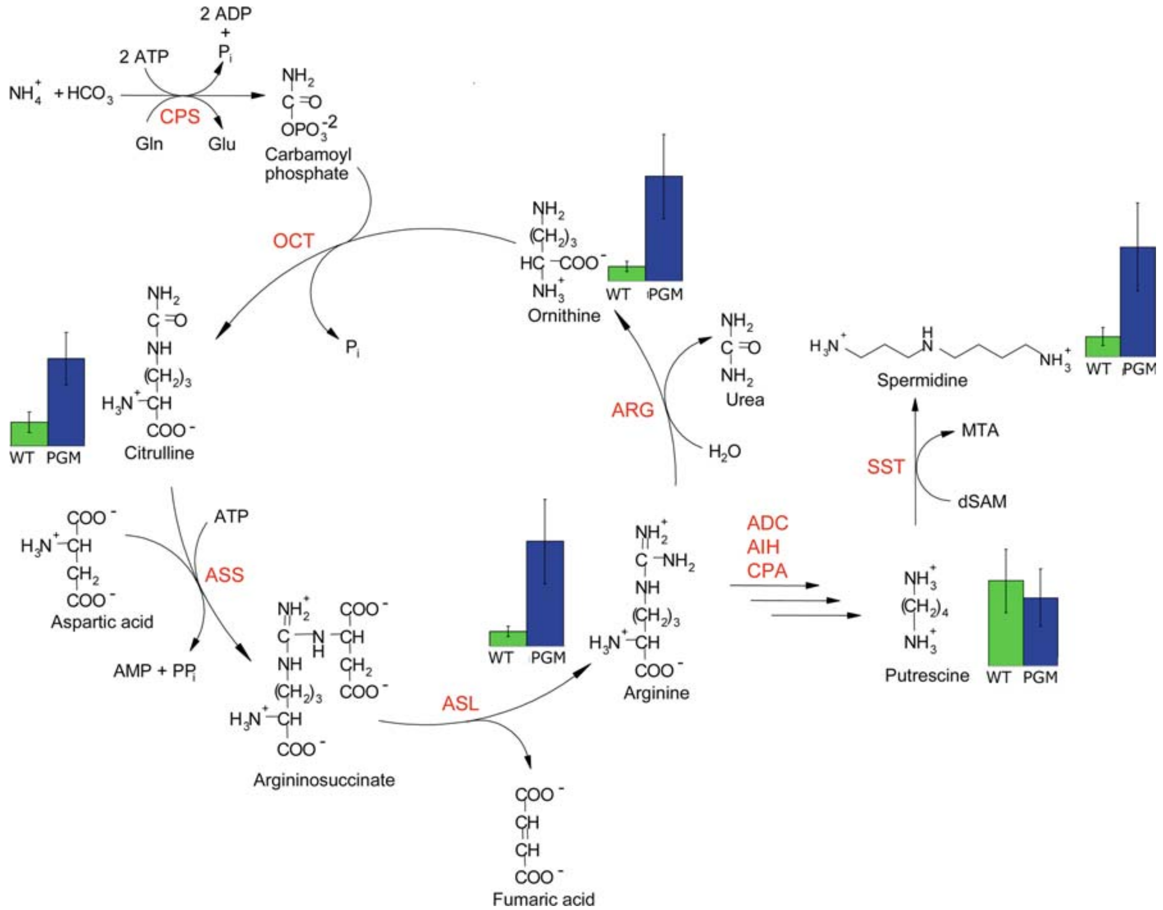

Figure 6. Urea cycle in *A. thaliana* as suggested by TAIR (http://www.arabidopsis.org/tools/aracyc/). The differences between WT and PGM plants are shown as average values (for discussion see text).

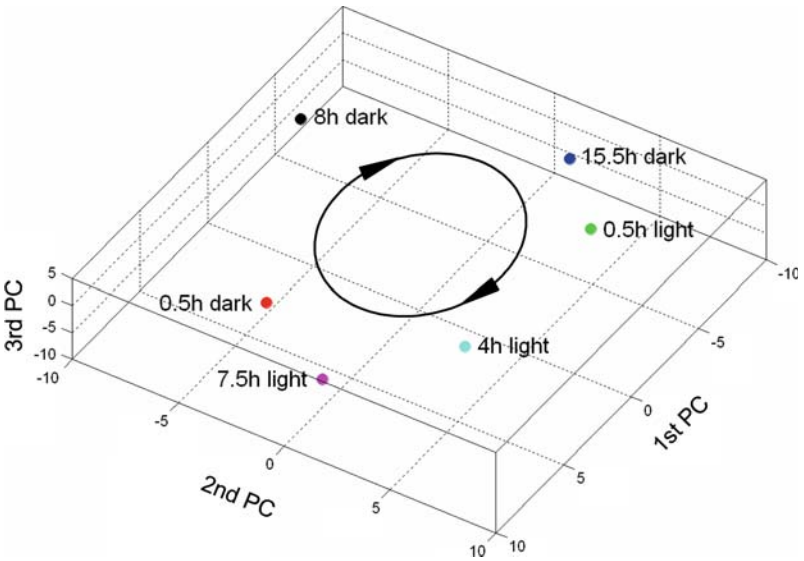

Figure 7. PCA of integrated metabolite–protein data of PGM plants during a diurnal cycle. The diurnal trajectory is clearly visible. Data points that are close neighbors were sampled at shorter time intervals (15.5 h dark and 0.5 h light; 7.5 h light and 0.5 h dark).

Yves Gibon, Oliver Blaesing, and Ralf Steuer for support and valuable discussion.

References

Blaschke, T. and Wiskott, L. (2004). CuBICA: independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Trans. Signal Process.* **52**, 1250–1256.

Boyes, D.C., Zayed, A.M., Ascenzi, R., McCaskill, A.J., Hoffman, N.E., Davis, K.R. and Gorlach, J. (2001). Growth stage-based phenotypic analysis of arabidopsis: a model for high throughput functional genomics in plants. *Plant Cell* **13**, 1499–1510.

Camacho, D., Fuente, A.D.L. and Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics* **1**, 53–63.

Cao, Y., Williams, D.D. and Williams, N.E. (1999). Data transformation and standardization in the multivariate analysis of river water quality. *Ecol. Appl.* **9**, 669–677.

Caspar, T., Huber, S.C. and Somerville, C. (1985). Alterations in growth, photosynthesis, and respiration in a starchless mutant of *Arabidopsis thaliana* (L) deficient in chloroplast phosphoglucomutase activity. *Plant Physiol.* **79**, 11–17.

Castrillo, J.O. and Oliver, S.G. (2004). Yeast as a touchstone in post-genomic research: strategies for integrative analysis in functional genomics. *J. Biochem. Mol. Biol.* **37**, 93–106.

Cichocki, A. and Amari, S. (2002). *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Wiley.

Collister, J.W., Rieley, G., Stern, B., Eglinton, G. and Fry, B. (1994). Compound-specific delta-C-13 analyses of leaf lipids from plants with differing carbon-dioxide metabolisms. *Organic Geochem.* **21**, 619–627.

Cooper, M., Chapman, S., Podlich, D. and Hammer, G. (2002). The GP problem: quantifying gene-to-phenotype relationships. *In Silico Biol.* **2**, 151–164.

Diamantaras, K. and Kung, S. (1996). *Principal Component Neural Networks*. Wiley, New York.

Fiehn, O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161.

Geiger, D.R. and Servaites, J.C. (1994). Diurnal regulation of photosynthetic carbon metabolism in C-3 plants. *Ann. Rev. Plant Physiol. – Plant Mol. Biol.* **45**, 235–256.

Gerhardt, R., Stitt, M. and Heldt, H.W. (1987). Subcellular metabolite levels in spinach leaves – regulation of sucrose synthesis during diurnal alterations in photosynthetic partitioning. *Plant Physiol.* **83**, 399–407.

Gibon, Y., Blaesing, O., Hannemann, J., Carillo, P., Hohne, M., Hendriks, J., Palacios, N., Cross, J., Selbig, J. and Stitt, M. (2004a). A Robot-based platform to measure multiple enzyme activities in *Arabidopsis* using a set of cycling assays: comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness. *Plant Cell* **16**, 3304–3325.

Gibon, Y., Blasing, O.E., Palacios-Rojas, N., Pankovic, D., Hendriks, J.H.M., Fisahn, J., Hohne, M., Gunther, M. and Stitt, M. (2004b). Adjustment of diurnal starch turnover to short days: depletion of sugar during the night leads to a temporary inhibition of carbohydrate utilization, accumulation of sugars and post-translational activation of ADP-glucose pyrophosphorylase in the following light period. *Plant J.* **39**, 847–862.

Glinski, M., Romeis, T., Witte, C., Wienkoop, S. and Weckwerth, W. (2003). Stable isotope labeling of phosphopeptides for multiparallel kinase target analysis and identification of phosphorylation sites. *Rapid Commun. Mass Spectrom.* **17**, 1579–1584.

Goodacre, R. (2003). Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vibrat. Spectroscopy* **32**, 33–45.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* **22**, 245–252.

Halket, J.M., Przyborowska, A., Stein, S.E., Mallard, W.G., Down, S. and Chalmers, R.A. (1999). Deconvolution gas chromatography mass spectrometry of urinary organic acids – potential for pattern recognition and automated identification of metabolic disorders. *Rapid Commun. Mass Spectrom.* **13**, 279–284.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis*. J. Wiley.

Ihmels, J., Levy, R. and Barkai, N. (2004). Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **22**, 86–92.

Kell, D.B. (2002). Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Rep.* **29**, 237–241.

Kell, D. and Mendes, P. (2000). Snapshots of systems: metabolic control analysis and biotechnology in the postgenomic era in Cornish-Bowden, A.J. and Cardenas, M.L. (Eds.) *Technological and Medical Implications of Metabolic Control Analysis*. Kluwer Academic Publishers, Netherland, pp. 2–25.

Kovats, E. (1958). Gas-Chromatographische Charakterisierung Organischer Verbindungen. 1. Retentionsindices Aliphatischer Halogenide, Alkohole, Aldehyde Und Ketone. *Helv. Chim. Acta* **41**, 1915–1932.

Leonard, C. and Sacks, R. (1999). Tunable-column selectivity and time-of-flight detection for high-speed GC/MS. *Anal. Chem.* **71**, 5177–5184.

Lippincott, J. and Apostol, I. (1999). Carbamylation of cysteine: a potential artifact in peptide mapping of hemoglobins in the presence of urea. *Anal. Biochem.* **267**, 57–64.

Nicholson, J.K., Connelly, J., Lindon, J.C. and Holmes, E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discovery* **1**, 153–161.

Nicholson, J.K., Lindon, J.C. and Holmes, E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29**, 1181–1189.

Ott, K.H., Aranibar, N., Singh, B.J. and Stockton, G.W. (2003). Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* **62**, 971–985.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. and Gygi, S.P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000). Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. *Plant J.* **23**, 131–142.

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

Roweis, S.T. and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326.

Sauter, H., Lauer, M. and Fritsch, H. (1991). Metabolic profiling of plants – a new diagnostic-technique. *Acs Symp. Series* **443**, 288–299.

Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. and Selbig, J. (2004). Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **20**, 2447–2454.

Scholz, M. and Vigário, R. (2002). Nonlinear PCA: a new hierarchical approach. *Proc. ESANN*. 439–444.

Steuer, R., Kurths, J., Fiehn, O. and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**, 1019–1026.

Sumner, L.W., Mendes, P. and Dixon, R.A. (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817–836.

Tabb, D.L., McDonald, W.H. and Yates, J.R. (2002). DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1**, 21–26.

ter Kuile, B.H. and Westerhoff, H.V. (2001). Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* **500**, 169–171.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y. and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**, 914–939.

Trethewey, R.N., Krotzky, A.J. and Willmitzer, L. (1999). Metabolic profiling: a Rosetta Stone for genomics?. *Curr. Opin. Plant Biol.* **2**, 83–85.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L. and Fernie, A.R. (2003). Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* **4**, 989–993.

Veriotti, T. and Sacks, R. (2001). High-speed GC and GC/time-of-flight MS of lemon and lime oil samples. *Anal. Chem.* **73**, 4395–4402.

Viant, M.R. (2003). Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem. Biophys. Res. Commun.* **310**, 943–948.

Wagner, A. (1997). Causality in complex systems. *Biol. Philos.* **14**, 83–101.

Watson, J.T., Schultz, G.A., Tecklenburg, R.E. and Allison, J. (1990). Renaissance of Gas-chromatography time-of-flight mass-spectrometry – meeting the challenge of capillary columns with a beam deflection instrument and time array detection. *J. Chromatogr.* **518**, 283–295.

Webb, J.W., Gates, S.C., Comiskey, J.P. and Weber, D.F. (1986). Metabolic profiling of corn plants using HPLC and GC/MS. Abstracts of Papers of the American Chemical Society 191, 70-ANYL.

Weckwerth, W. (2003). Metabolomics in systems biology. *Ann. Rev. Plant Biol.* **54**, 669–689.

Weckwerth, W. and Fiehn, O. (2002). Can we discover novel pathways using metabolomic analysis? *Curr. Opin. Biotechnol.* **13**, 156–160.

Weckwerth, W., Loureiro, M.E., Wenzel, K. and Fiehn, O. (2004a). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. USA* **101**, 7809–7814.

Weckwerth, W., Tolstikov, V. and Fiehn, O. (2001). Metabolomic characterization of transgenic potato plants using GC/TOF and LC-MS analysis reveals silent metabolic phenotypes. *Proceedings of the 49th ASMS Conference on Mass spectrometry and Allied Topics*, American Society of Mass Spectrometry, Chicago, pp. 1–2.

Weckwerth, W., Wenzel, K. and Fiehn, O. (2004b). Process for the integrated extraction identification, and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* **4**, 78–83.

Wienkoop, S., Glinski, M., Tanaka, N., Tolstikov, V., Fiehn, O. and Weckwerth, W. (2004a). Linking protein fractionation with multidimensional monolithic RP peptide chromatography/mass spectrometry enhances protein identification from complex mixtures even in the presence of abundant proteins. *Rapid Commun. Mass Spectrom.* **18**, 643–650.

Wienkoop, S., Zoeller, D., Ebert, B., Simon-Rosin, U., Fisahn, J., Glinski, M. and Weckwerth, W. (2004b). Cell-specific protein profiling in *Arabidopsis thaliana* trichomes: identification of trichome-located proteins involved in sulfur metabolism and detoxification. *Phytochemistry* **65**, 1641–1649.

Winter, H. and Huber, S.C. (2000). Regulation of sucrose metabolism in higher plants: localization and regulation of activity of key enzymes. *Crit. Rev. Biochem. Mol. Biol.* **35**, 253–289.