II.2 Chemometrics in Metabolomics – An Introduction

J. Trygg[1], J. Gullberg[2], A.I. Johansson[2], P. Jonsson[1], and T. Moritz[2]

1 Introduction

In the post-genomics era, the use of methodologies that enable transcriptomic, proteomic and metabolomic data to be analysed in detail have revolutionized biological investigations. One of the major advantages with metabolomics investigations compared to traditional target metabolite analysis is that metabolomics data can give an unbiased view of changes in metabolism during environmental, genetic or developmental changes. Instead of tracking only a few metabolites, changes in relative amounts in 300 to 1000 or even more metabolites can be recorded and analysed, covering all major metabolic pathways. This development has accentuated the need to apply and further develop multivariate methodology. Chemometrics (see Eriksson et al. 2001) provides tools to make good use of measured data, enabling practitioners to make sense of measurements and to model quantitatively and produce visual representations of information. Today, chemometrics has grown into a well established data analysis tool in areas such as multivariate calibration, quantitative structure-activity modeling, pattern recognition and multivariate statistical process monitoring and control. Although seemingly diverse disciplines, the common denominators in these applications are that high complexity data tables are generated and that these data tables can be analysed and interpreted by means of chemometric methods.

In chemometrics, there are three basic categories of analysis (Fig. 1):

1. Exploratory analysis (Fig. 1A). This gives an overview of all the data in order to detect trends, patterns or clusters.
2. Classification analysis and discriminant analysis (Fig. 1B), which classifies samples into categories or classes, for example wild-type and mutant.
3. Regression analysis and prediction models (Fig. 1C) are used when a quantitative relationship between two blocks of data is sought. For example, when prediction of growth or fiber properties from mass spectrometry data.

However, in biology, chemometric methodology has still been largely overlooked in favour of traditional statistics. It is not until recently that the

[1] Research Group for Chemometrics; Organic Chemistry, Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden, e-mail: johan.trygg@chem.umu.se
[2] Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden, e-mail: thomas.moritz@genfys.slu.se

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

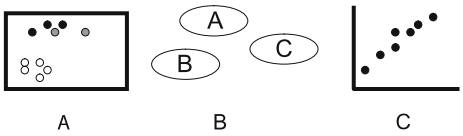

Fig. 1. Overview of the basic categories of chemometrics analysis: **A** overview of data structure; **B** classification and discriminant analysis; **C** regression analysis

overwhelming size and complexity of the 'omics' technologies has driven biology towards the adoption of chemometric methods. Here we will give an introduction to chemometrics and also give examples of why and when chemometrical methodologies should be used.

2 Theory and Methods

2.1 Making Data Contain Information – Design of Experiments

In experimental biology, e. g. when investigating how a number of different environmental factors (e. g. temperature, day length, nutrition) affect different responses such as growth, transcript profiles and metabolite profiles in plants, there is a need to carry out experiments in a systematic way. One way to investigate how the factors affect the plant's responses is to Change One Factor at a Time, i. e. the COST approach. This approach has severe problems: (1) finding optimal conditions for experiments (e. g. method development), (2) unnecessarily many experiments are needed (inefficiency), (3) ignores interaction among variables (lost information) and (4) provides no map over the experimental space.

Design of Experiments (DOE) (Lundstedt et al. 1998) is the methodology of how to conduct and plan experiments in order to extract the maximum amount of information in the fewest number of runs. The basic idea is to devise a small set of experiments, in which all pertinent factors are varied systematically. It is a fundamental tool for planning experiments and making data informative by simultaneously, albeit in a structured way, varying controllable factors (e. g. environmental conditions, instrument settings, experimental procedures) of the studied system. Today they comprise a tool box for virtually any experimental problem.

2.1.1 Stages in the DOE Process

Most of us can only grasp the effect of one factor at a time in our minds, and that often leads us into the inefficient COST approach. We need the mathematics (and the computer) to keep track of the factors and their combinations.

In summary, (1) all factors are varied together over a set of experimental runs, (2) noise is decreased by means of averaging, (3) the functional space is efficiently mapped, interactions and synergisms are seen.

1. What do I want? – formulate question(s) stating the objectives and goals of the investigation. For example identify factors (e. g. temperature, day length, nutrition) and factor ranges (e. g. 15–25 °C, 6–12 h, 1–10 mmol N/L) that affects flowering time.
2. Screening design – finding out a little about many factors. Which factors are the dominating ones in controlling flowering time? Screening designs provide simple models with information about dominating variables, and information about ranges. Pareto's principle states that 20% of the data (factors) account for 80% of the information. Different types of screening designs exist – which one to choose depends on the problem. The most common one is the fractional factorials design (Fig. 2). The full factorial design is a set of experimental runs where every level of a factor is investigated at both levels of all the other factors. It requires $N = 2^k$ number of runs for k factors. Investigating more than five factors with the full factorial design can in some cases become time consuming, i. e. $2^5 = 32, 2^6 = 64, 2^7 = 128$ experiments, etc. Instead, performing a *fractional factorial design* reduces

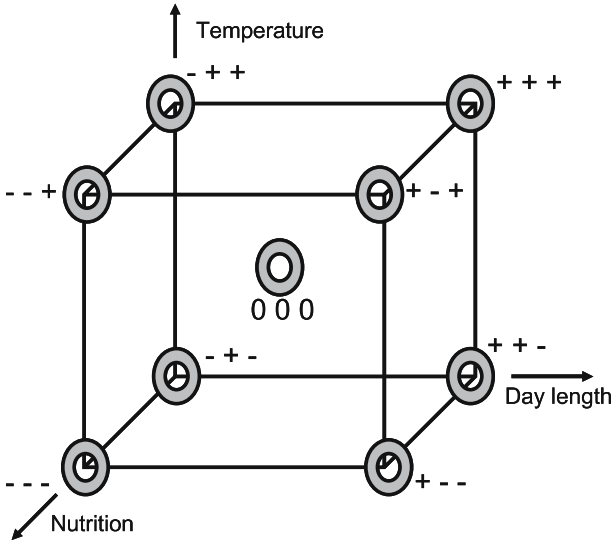

Fig. 2. Example of a full factorial design of experiments (DOE) for investigating how three factors (temperature, day length and nutrition) control flowering time. Varying the three factors at two levels (coded as +/-) requires $2^3 = 8$ experiments + center points. Each experiment according to the design set of experiments is *marked with a circle* in the figure. Evaluating the results from such an experimental design reveals the influence of each of the different factors separately and also any interactions between them. DOE is the only feasible approach to separate cause and effect from each other

that number quickly without the loss of too much information regarding the estimation of factors involved. Fractional factorial design takes advantage of the fact that three-way and higher interactions are seldom significant. It requires only $N = 2^{k-p}$ number of runs for k factors, where p is set manually. For example five factors can be run in only $2^{5-2} = 8$ experiments instead of $2^5 = 32$ experiments compared to the full factorial design. Fractional factorial design takes advantage of the fact that three-way and higher interactions are seldom significant. The downside, of course, for not performing all experiments, is that confounding patterns are present. In other words, the estimated effects are not "pure" but instead mixed with higher degree interaction effects. This loss of information is the prize we need to pay for the reduction of the number of experiments. The degree of confounding is determined by the choice of p.

3. Response surface modeling (RSM) and optimization (few factors) – after screening the factors involved in, e. g. determination of flowering time or derivatization of metabolites, the goal of the investigation is usually to create a valid map of the experimental domain (local space) given by the significant factors and their ranges. This is done with a quadratic polynomial model. The higher order models have an increased complexity, and therefore also require more experiments/factors than screening designs. Different types of RSM designs include Central composite designs, Box Behnken designs and D-optimal designs (see, e. g. Lundstedt et al. 1998 for more information).
4. Robustness testing – in robustness testing of, for instance, an analytical method, the aim is to explore how sensitive the responses are to small changes in the factor settings, e. g. temperature. Ideally, a robustness test should show that the responses are not sensitive to small fluctuations in the factors, that is, the results are the same for all experiments. Robustness testing is usually applied as the last test just before the release of a product or a method. The fractional factorial design is usually applied here.

Plant metabolomic studies typically constitute a set of samples from Arabidopsis wild types and mutants. Assume that these have been subjected to different external conditions such as variation in day length and temperature. Design of Experiments can then be used to select representative samples, related to the biological question we are investigating (how flowering time is affected by temperature, day length, nutrition). An experimental design in three factors can be setup, with factor 1 (temperature), factor 2 (day length), and factor 3 (nutrition). In total, only eight different experiments equal 2^k where $k = 3$ factors are required to explore the experimental space. In addition, a number of replicates, typically three experiments, are added to estimate the noise level. By adding extra experiments, one can investigate more thoroughly the day length and temperature dependence (increase the number of different day lengths and temperatures).

2.2 The Data Table, X-matrix

In plant metabolomics studies, typically a set of samples are characterised using modern instrumentation such as GC/MS, LC/MS or H^1-NMR spectroscopy. The choice of instrument (see Sumner et al. 2003; Dunn et al. 2005) and experimental procedure (Gullberg et al. 2004) are important and largely determined by the biological system and the scientific question. Design of Experiments can here be used to optimize the experimental protocol.

In contrast to a ^{1}H-NMR spectrum, GC/MS and LC/MS data must be processed before multivariate analysis. The reason is the two-dimensional nature (chromatogram/mass spectra) of the data for each sample. For GC/MS data, curve resolution or deconvolution methods are mainly applied for data processing (see, e. g. Halket et al. 1999; Jonsson et al. 2005a). This gives a resolved spectral and chromatographic profile for each detected compound. The 1D multivariate profile used to characterize each sample is made up of the integrated areas of all detected chromatographic peaks. The corresponding mass spectrum and retention index are used for identification purposes (Schauer et al. 2005). For LC/MS data, curve resolution can be applied (e. g. Idborg-Björkman et al. 2003) or a peak detection algorithm that identifies all chromatographic peaks and uses their integrated areas as the multivariate profile characterizing that sample (e. g. Andreev et al. 2003). Another alternative is to sum the chromatographic direction to create a 1D multivariate profile produced by the total intensity over all mass spectral channels (e. g. Allen et al. 2004). Recently, partly alternative methodologies have been applied to GC/MS data (Jonsson et al. 2004, 2005a) and LC/MS data (Jonsson et al. 2005b) where all samples are processed simultaneously and a common set of descriptor variables are extracted.

After, e. g. the GC/MS analysis, we now have a multivariate profile (300–1000 s of variables) for each sample that is a fingerprint of the inherent properties (e. g. phenotype) for each sample. For multiple samples we can therefore construct a two-dimensional data table, an X matrix, by stacking each sample on top of each other. The question is then, how do we go about analysing this multivariate, highly collinear and complex data set? The univariate approach (e. g. student's t-test [Jackson 1991]) is not recommended. It assumes independent variables in X (i. e. more samples than variables) and this creates problems with interpretation, spurious correlations (so called Type I, II errors) and the evident risk of missing information in combinations of variables. Traditional statistical methods (e. g. multiple linear regression, MLR) are also not recommended. They also assume independent variables and have difficulties with noisy data (Eriksson et al. 2001). Instead, multivariate analyses based on projection methods represent a number of efficient and useful methods for the analysis and modeling of these complex data. Projection methods convert the multi-dimensional data table into a low-dimensional model plane, usually consisting of two to five dimensions. Principal component analysis (PCA) (Jackson 1991) and partial least squares (PLS) (Wold et al. 1984) methods are

two widely used methods that can handle incomplete, noisy and collinear data structures.

2.3 Geometrical Interpretation of a Data Table

An easy way to understand and appreciate projection based methods is to translate the data table into a swarm of points in a multi-dimensional space. For a data table or matrix X, with N rows (biological samples) and K columns (e. g. relative amounts of different metabolites), each row (individual sample) can be represented as a point in a K-dimensional space. Its position in this space is given by its coordinates, i. e. its values in each of the K columns. Repeating this for all N rows in a matrix, we have produced a swarm of points in K-dimensional space. Points (samples) that lie close to each other in this multi-dimensional space are more biologically similar to each other than points that lie far apart (dissimilar). Projection methods find a model hyperplanes of much lower dimensionality that closely approximates X, i. e. the swarm of points. Figure 3 gives an overview of how multivariate projection methods work.

2.4 Principal Component Analysis

Principal Component Analysis (PCA) is the workhorse in chemometrics. It is a multivariate projection method designed to extract and display the systematic variation in a data matrix X. The first two *principal components* define a plane, a window into the K-dimensional space. By projecting each of the sample points (in K-dimensional space) onto this two-dimensional sub-space, it is possible to visualize all the samples. The coordinates of each of these samples projected onto this plane are called *scores* T, and they are weighted averages of all X-variables (e. g. metabolites). Hence the visualization of these scores T is called a *score plot*. The score plot is very informative because it gives an overview of all samples in X and how they relate to each other. It may reveal groupings of samples (clusters), trends and outliers (deviating samples). e. g. two genotypes (wild type and mutant) would show up as two distinct clusters of samples, representing wild type and mutant samples respectively. In addition, an experiment that suffered from a broken GC-vial would translate into an unique point in the score plot, i. e. an outlier (Fig. 3).

The score plot allows us to investigate the relation among the samples, but once interesting patterns are found (groupings, outliers etc.), it is possible to understand the reason for this, i. e. what variables (e. g. metabolites) are responsible for this pattern found in the score plot. Hence, there also exists a corresponding plot related to the measured variables (metabolites), i. e. the columns in the X matrix. This plot is known as the *loading plot* P and describes the influence (weight) of the X-variables (metabolites) in the model. An important feature is that directions in the score plot correspond to directions in the

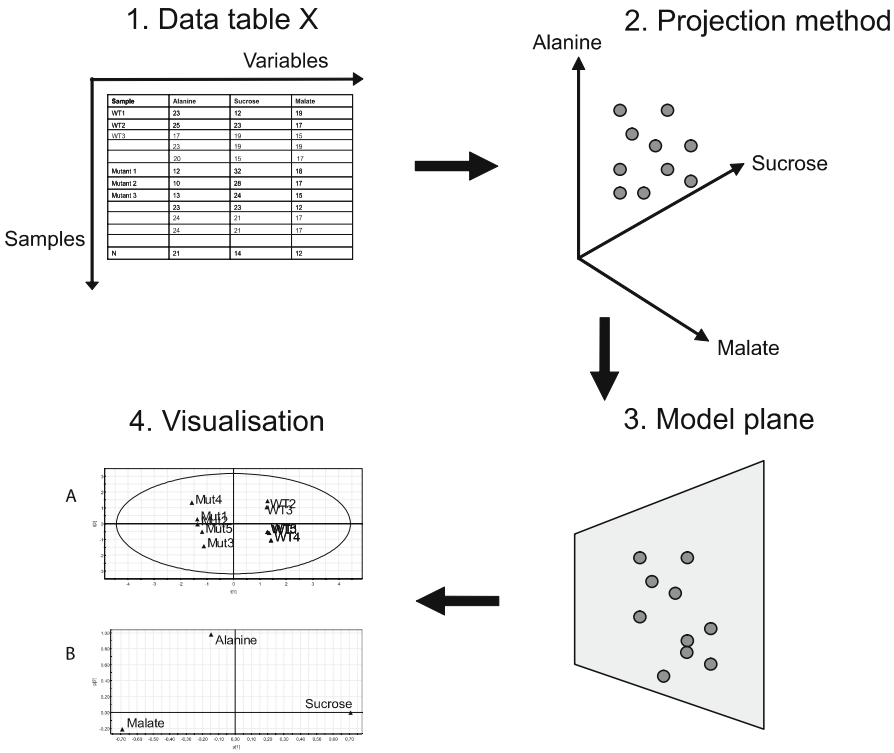

Fig. 3. (1) Each row (representing one biological sample) in a data table with K = 3 variables can be represented as one point in a K = 3 dimensional space. The position of that point is given by the coordinates given by the values in each of the K = 3 variables. (2) Repeating this for all rows (samples) in a data table produces a swarm of points in K = 3 dimensional space. Points (samples) that are close to each other have more similar biological properties than points that are far apart. (3) Projection methods such as PCA, finds a representative low-dimensional plane (here two-dimensional) that is a good summary of the variation in the X data table (swarm of points). (4) This model plane can then be visualised in scatter plots (A) and provides an overview, e. g. if there are any groupings, trends or outliers in the data. For example in the figure (A) there is a clear separation between the *Arabidopsis* wild type and mutant. It is also possible to understand the reason for this separation by looking at the direction of the model plane with respect to the original axes (original variables). These are summarized in the PCA model loadings, P (B)

loading plot (Fig. 3). This is a powerful tool for understanding the underlying patterns in the data.

The PCA model can be expressed as

$$\text{Model of X: } X = TP^T + E$$

where T are the scores, P defines the loadings, and E represent the residual matrix. The residual matrix E contains the residuals for each sample between

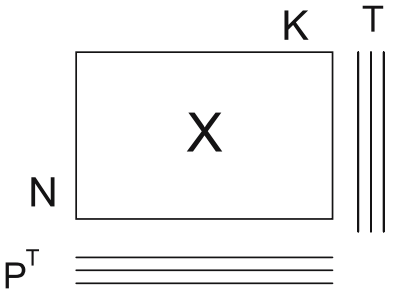

Fig. 4. PCA summarise all variation in X into a few new variables called scores T. These new variables are linearly weighted combinations of the original X-variables. The loadings P contain the weights used for each X-variable and thus reveal the influence of individual X-variables

its point in K-dimensional space and its point on the model plane. The residuals are important for detection of outliers and for defining the model boundaries (see Fig. 4).

2.5 Partial Least Squares Projections to Latent Structures (PLS)

The PLS method is used instead of the PCA method when additional knowledge about each sample exists, the Y matrix, e. g. genotype of each sample (wild type/mutant). The sample information according to the design matrix from the Design of Experiments (see Sect. 2.1) is often used as a Y matrix. Hence, PLS represents the regression analogy of PCA working with two matrices, X and Y (Wold et al. 1984). It is one of the most common methods when a quantitative relationship between a descriptor matrix X and a response matrix Y is sought. The Y matrix can contain both quantitative (e. g. glucose concentration) and qualitative (genotype) information. This additional sample information in Y is used by the PLS method to focus the model plane to capture the *Y-related variation* in X, e. g. separation between genotypes, rather than providing an overall view of *all variation* in the data as done by the PCA model. In addition, the PLS method can also be used to predict the properties (Y-values) of new unknown samples, e. g. predict the glucose concentration or genotype.

The Y matrix consists of the same number of rows as the X matrix. Each column in Y indicate a certain property, e. g. glucose concentration or genotype for each sample. When Y contains qualitative information such as genotype, the number of columns in Y equals the number of classes. Each row in Y describes the group membership for that sample where "1" indicates class belonging for that sample and "0" does not. When Y is qualitative, the PLS method is called PLS Discriminant Analysis (PLS-DA), to distinguish it from the situation when Y is quantitative.

3 Example: Metabolomics Study on Arabidopsis Mutants

We will work through a metabolomics example using GC/MS data from the analysis of *Arabidopsis* extracts. Shoots of higher plants are characterized by axillary branching, where the shoot branches develop from shoot meristems located between a leaf and the shoot stem. The control of axillary shoot growth (branching) is not well understood, but it is known that several internal factors such as the plant hormones IAA and cytokinins are involved (McSteen and Leyser 2005). Mutations screens in *Arabidopsis* have identified four loci involved in the repression of axillary bud growth, *MAX1–4*. Based on the mutants, it is now suggested that an unknown transmittable substance might be involved in controlling branching (see McSteen and Leyser 2005). The biosynthesis of this compound in *Arabidopsis* is catalyzed by a number of MAX (more-axillary growth) proteins.

We have used a metabolomics approach to classify and identify the metabolic differences between the MAX-mutants. Root samples from WT, max3 and max4 mutants were analysed by GC/TOFMS as described by Gullberg et al. (2004). The GC/MS data was processed by hierarchical multivariate curve resolution (Jonsson et al. 2005a), and the obtained X-matrix was thereafter subjected to PCA and PLS-DA analysis. The GC/MS processing resulted in 514 resolved peak areas. Log transformation, column centering and scaling to unit variance was done on the resolved peak areas (X-matrix) prior to modeling and two dummy Y-variables were constructed based on the class belonging of each sample to the

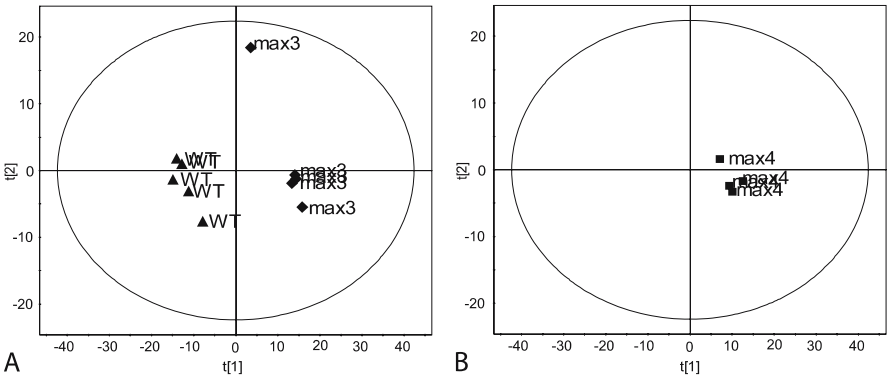

Fig. 5. A PLS-DA score-plot from the analysis of metabolite profiles in roots of *Arabidopsis* WT, max3 and max4. The PLS-DA model is based on WT and max3. The X-matrix was centered and scaled to unit variance. The explained variation in the X-matrix (R^2X) is 0.74, the explained variation in the Y matrix (R^2Y) is 0.99 and the predictive ability according to sevenfold cross-validation (Q2) is 0.84. R^2X is the cumulative modelled variation in X, R^2Y is the cumulative modelled variation in Y and Q^2Y is the cumulative predicted variation in Y, according to cross-validation. The range of these parameters is 0–1, where 1 indicates a perfect fit. **B** Based on the model max4 samples were predicted into the model showing that max3 and max4 are very similar regarding metabolic content (compare position score plot in **A**)

genotypes, WT and max3. The PLS-DA model score plot is shown in Fig. 5A. The score plot reveals the relationship among the samples. It is clear from the figure that the model plane displays a clear separation of the two genotypes.

To validate the model results, predictions were made for the genotype max4, using the calculated PLS-DA model based on the other sample-set (WT and max3). The results, shown in the obtained PLS-DA score plot (Fig. 5B) predicted that the max4 is closer to max3 than WT. This is consistent with the facts that max3 is very similar to the max4 genotype, where the MAX3 and MAX4 proteins use the same substrate (Schwarz et al. 2005). Interpretation of the first weight vector (w1) from the PLS-DA model, as described by Trygg and Wold (2002), together with the 99% confidence intervals calculated using jack-knifing (Martens and Martens 2000), highlighted 64 significant variables (metabolites) differing between WT and max4. The importance of these metabolites is a part of *biological validation* of the data set. The *statistical validation* was done by prediction of the max3 mutants into the WT/max4 model. Both type of validation is of importance for validating the multivariate data set.

4 Summary and Future Prospectives

Multivariate projection methods, e. g. PCA and PLS, represent a useful and versatile technology to modelling, monitoring and prediction of complex problems and data structures encountered within metabolomics and other 'omics' disciplines. The common denominator is that high complexity data tables are generated and that these data tables can be analysed and interpreted by means of chemometric methods. The principal component analysis (PCA) method summarizes the variation in a data table X into a model plane (the scores T). A scatter plot of these scores gives an overview of the samples (observations) and how they relate to each other, e. g. if there are groupings or trends or deviating samples and so on. In order to interpret the patterns found in a score plot one examines the corresponding loading plot (P). The loadings P reveal how each variable contributes to the separation among samples in the model plane and also gives insights into the relative importance of each variable.

However, one fundamental property is that the data does contain relevant information regarding our biological question. In other words, how to maximise the information content in the data? The traditional way to Change One Factor at a Time, i. e. the COST approach, is not recommended. Design of Experiments (DOE) is the methodology of how to conduct and plan experiments in order to maximize information in the data in the fewest number of runs. A proper experimental design will reveal the influence of each of the different factors separately and also any interactions between them. DOE is the only feasible approach to separate cause and effect from each other. Therefore is DOE in combination with chemometrical analysis a powerful way of planning, conducting and evaluating metabolomics experiments.

One common discussion point in the analysis of "omics" data is how to correlate several types of data, usually with different data structures. Systems biology seeks to integrate information from multiple parts of a biological system in a holistic attempt to understand the whole system. There are still many obstacles and hurdles to overcome in order to succeed. One of these relates to how the actual integration of the different types of data will be done. Hence, the advancement of systems biology depends heavily on the ability to integrate multiple profiling techniques (e. g. transcriptomics, proteomics, GC/MS, LC-NMR). The current multivariate statistical methods (e. g. the PLS method) lacks the proper model structure to describe these types of data structures, because they focus only on the *correlation pattern* among multiple data tables (e. g. X = microarrays vs Y = metabolomics data) and not on the *non-correlated variation* among these data tables which, in a biological sense, can be of equal interest. It has also been demonstrated that, because of this, the interpretation of these models are negatively affected (Trygg and Wold 2002), e. g. positive correlation patterns are interpreted as negligible or even flipped and become negative. This is a fundamental problem as we certainly cannot expect that all variation in transcript and metabolite levels co-vary. Fortunately, recent advances in chemometrics provide the ability to compare multiple data sets with each other. Novel extensions of the PLS method, called O-PLS (Trygg and Wold 2002) and O2-PLS (Trygg 2002) contain the model structure to support both these features. In addition, the O2-PLS method is bi-directional which means that the flow of information can go in both ways, from X (e. g. microarray) to Y (e. g. metabolomics) and vice versa. Hence, the O2-PLS methodology will be important in selecting what genes or metabolites are important to do further experimentation upon, e. g. understanding biomarker patterns and selecting genes for knockout studies. The O2-PLS methodology can also be extended to more than two data tables, hence it nicely fits into the framework of a combined profiling approach.

Acknowledgements. The Swedish Research Council, Wallenberg Consortium North (WCN), the Kempe foundation, EU strategic funding, Knut and Alice Wallenberg Foundation (JT) and Strategic Research Funding (SSF) are acknowledged for financial support. Professor Ottoline Leyser, York, UK, for allowing us to show data from the max-mutant project, and Dr. Miyako Kusano, RIKEN Plant Science Centre, Yokohama, Japan for the initial analysis of metabolites in the max-mutants.

References

Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG, Kell DB (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. Nature Biotechnol 21:692–696

Andreev VP, Rejtar T, Chen HS, Moskovets EV, Ivanov AR, Karger BL (2003) A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. Anal Chem 75:6314–6326

Dunn WB, Bailey NJC, Johnson HE (2005) Measuring the metabolome: current analytical technologies. Analyst 130:606–625

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multi and megavariate data analysis. Umetrics (www.umetrics.com), ISBN 91–973730-1-X

Gullberg J, Jonsson P, Nordström A, Sjöström, M, Moritz T (2004) Optimisation of preparation of plant samples for metabolic profiling by GC-MS. Anal Biochem 331:283–295

Halket JM, Przyborowska A, Stein SE, Mallard WG, Down S, Chalmers RA (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

Idborg-Björkman H, Edlund, PO, Kvalheim OM, Schuppe-Koistinen I, Jacobsson SP (2003) Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. Anal Chem 75:4784–4792

Jackson JE (1991) A users guide to principal components. Wiley, New York

Jonsson P, Gullberg J, Nordström A, Kowalczyk M, Sjöström M, Moritz T (2004) A strategy for extracting information from large series of non-processed complex GC/MS data. Anal Chem 76:1738–1745

Jonsson P, Johansson AI, Gullberg J, Trygg J, A J, Grung B, Marklund S, Sjöström M, Antti H, Moritz T (2005a) Highthroughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. Anal Chem 77:5635–5642

Jonsson P, Bruce SJ, Moritz T, Trygg J, Sjostrom M, Plumb R, Granger J, Maibaum E, Nicholson JK, Holmes E, Antti H (2005b) Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. Analyst 130:701–707

Lundstedt T, Seifert E, Abramo L, Thelin B, Nyström A, Pettersen J, Bergman R (1998) Experimental design and optimization. Chem Intel Lab Systems 42:3–40

Martens H, Martens M (2000) Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Qual Pref 11:5–16

McSteen P, Leyser O (2005) Shoot branching. Annu Rev Plant Biol 56:353–374

Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L et al (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett 579:1332–1337

Schwartz S, Qin XQ, Loewen MC (2005) The biochemical characterization of two Carotenoid cleavage enzymes from *Arabidopsis* indicates that a carotenoid-derived compound inhibits lateral branching. J Biol Chem 279:46940–46945

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

Trygg J (2002) O2-PLS for qualitative and quantitative analysis in multivariate calibration. J Chemometr 16:283–293

Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemometrics 16:119–128

Wold S, Ruhe A, Wold H, Dunn WJ III (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Statist Comput 5:735–743