Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models

Susanne Wiklund,[†] Erik Johansson,[‡] Lina Sjöström,[†] Ewa J. Mellerowicz,[§] Ulf Edlund,[†] John P. Shockcor,[||] Johan Gottfries,[†,⊥] Thomas Moritz,[§] and Johan Trygg*[,†]

Department of Chemistry, Umeå University, SE 901 87 Umeå, Sweden, Umetrics, Tvistevägen 48, P.O. Box 7960, SE 90719, Umeå, Sweden, Umeå Plant Science Center, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden, and AstraZeneca R&D, SE 43183 Mölndal, Sweden

Metabolomics studies generate increasingly complex data tables, which are hard to summarize and visualize without appropriate tools. The use of chemometrics tools, e.g., principal component analysis (PCA), partial least-squares to latent structures (PLS), and orthogonal PLS (OPLS), is therefore of great importance as these include efficient, validated, and robust methods for modeling information-rich chemical and biological data. Here the S-plot is proposed as a tool for visualization and interpretation of multivariate classification models, e.g., OPLS discriminate analysis, having two or more classes. The S-plot visualizes both the covariance and correlation between the metabolites and the modeled class designation. Thereby the S-plot helps identifying statistically significant and potentially biochemically significant metabolites, based both on contributions to the model and their reliability. An extension of the S-plot, the SUS-plot (shared and unique structure), is applied to compare the outcome of multiple classification models compared to a common reference, e.g., control. The used example is a gas chromatography coupled mass spectroscopy based metabolomics study in plant biology where two different transgenic poplar lines are compared to wild type. By using OPLS, an improved visualization and discrimination of interesting metabolites could be demonstrated.

Systems biology is a rapidly growing research field that aims to study the interactions between different components of a biological system with a holistic perspective. The goal is to understand how these interactions give rise to the function and behavior of the studied system, e.g., how gene regulation affects enzymes and metabolites in a metabolic pathway.[1,2] The study of low molecular weight molecules is usually referred to as metabonomics or metabolomics and has proven an important area of systems biology. Typically, in metabolomics studies, the aim is to pinpoint the putative metabolites that are related to drug toxicity, disease, and environmental or genetic variation.[3−10] However, metabolomics generate complex data tables that are hard to summarize and interpret without appropriate statistical and visualization tools. The use of chemometrics tools, e.g., principal component analysis (PCA),[11] partial least-squares to latent structures (PLS),[12,13] and orthogonal PLS (OPLS),[14] are therefore of great importance as they include efficient and robust methods for modeling, analysis, and interpretation of complex chemical and biological data. The results from a PCA model will indicate systematic trends within the data, i.e., clustering, time trends, etc. However, instrumental drift, artifacts, and other experimental variation will on occasions divert the focus of a PCA model to the systematic variation unrelated to the scientific question of interest. In such cases, there is a need for methods that make use of any

[*] To whom correspondence should be addressed. Phone:+46907866917. Fax: +4690138885. E-mail: Johan.Trygg@chem.umu.se.

[†] Umeå University.

[‡] Umetrics.

[§] Swedish University of Agricultural Sciences.

[||] Visiting fellow: Department of Biochemistry, University of Cambridge, Cambridge, UK.

[⊥] AstraZeneca R&D.

(1) Goodacre, R.; Vaidyanathan, S.; Dunn, W. B.; Harrigan, G. G.; Kell, D. B. *Trends Biotechnol.* **2004**, *22* (5), 245−252.

(2) Hollywood, K.; Brison, D. R.; Goodacre, R. *Proteomics* **2006**, *6*, 4716−4723.

(3) Marchesi, J. R.; Holmes, E.; Khan, F.; Kochhar, S.; Scanlan, P.; Shanahan, F.; Wilson, I. D.; Wang, Y. L. *J. Proteome Res.* **2007**, *6* (2), 546−551.

(4) Wang, C.; Kong, H. W.; Guan, Y. F.; Yang, J.; Gu, J. R.; Yang, S. L.; Xu, G. W. *Anal. Chem.* **2005**, *77* (13), 4108−4116.

(5) Wiklund, S.; Karlsson, M.; Antti, H.; Johnels, D.; Sjöström, M.; Wingsle, G.; Edlund, U. *Plant Biotechnol. J.* **2005**, *3* (3), 353−362.

(6) Catchpole, G. S.; Beckmann, M.; Enot, D. P.; Mondhe, M.; Zywicki, B.; Taylor, J.; Hardy, N.; Smith, A.; King, R. D.; Kell, D. B.; Fiehn, O.; Draper, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (40), 14458−14462.

(7) Le Gall, G.; Colquhoun, I. J.; Davis, A. L.; Collins, G. J.; Verhoeyen, M. E. *J. Agric. Food Chem.* **2003**, *51* (9), 2447−2456.

(8) Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Keun, H. C.; Craig, A.; Pearce, J. T. M.; Bruce, S. J.; Hardy, N.; Sansone, S. A.; Antti, H.; Jonsson, P.; Daykin, C.; Navarange, M.; Beger, R. D.; Verheij, E. R.; Amberg, A.; Baunsgaard, D.; Cantor, G. H.; Lehman-McKeeman, L.; Earll, M.; Wold, S.; Johansson, E.; Haselden, J. N.; Kramer, K.; Thomas, C.; Lindberg, J.; Schuppe-Koistinen, I.; Wilson, I. D.; Reily, M. D.; Robertson, D. G.; Senn, H.; Krotzky, A.; Kochhar, S.; Powell, J.; van der Ouderaa, F.; Plumb, R.; Schaefer, H.; Spraul, M. *Nat. Biotechnol.* **2005**, *23* (7), 833−838.

(9) Antti, H.; Ebbels, T. M. D.; Keun, H. C.; Bollard, M. E.; Beckonert, O.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Chemom. Intell. Lab. Syst.* **2004**, *73* (1), 139−149.

(10) Eriksson, L.; Antti, H.; Gottfries, J.; Holmes, E.; Johansson, E.; Lindgren, F.; Long, I.; Lundstedt, T.; Trygg, J.; Wold, S. *Anal. Bioanal. Chem.* **2004**, *380* (3), 419−429.

(11) Jackson, J. E. *A Users Guide to Principal Components*; John Wiley: New York, 1991.

(12) Wold, S.; Trygg, J.; Berglund, A.; Antti, H. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 131−150.

(13) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III. *SIAM J. Sci. Stat. Comput.* **1984**, *5* (3), 735−743.

(14) Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16* (3), 119−128.

10.1021/ac0713510 CCC: $40.75 © 2008 American Chemical Society
Published on Web 11/21/2007

a priori information to refocus the analysis toward the studied objectives by use of, for example, PLS or OPLS. This a priori information, e.g., object membership, constitutes an extra data table, Y, where the columns indicate sample information, which can be a discrete or continuous value. The advantage with OPLS compared to PLS is that the model is rotated[15] so that class separation is found in the first predictive component, t_p, also referred to as the correlated variation, and variation not related to class separation is seen in orthogonal components, also referred to as the uncorrelated variation. This separation of predictive and orthogonal components facilitates model interpretation.

In the present study, a new and efficient approach is presented for modeling of two or more classes. The paper will unravel OPLS discriminate analysis (OPLS-DA) model interpretation in multiclass situations with the following focus: (i) extraction of statistically and potentially biochemically significant metabolites by use of a S-plot and (ii) identification of shared and unique structures by the use of a shared and unique structure (SUS)-plot.

Metabolomics data normally comprise large dynamic ranges in metabolite concentration. Analysis of such multivariate data requires methodology that can handle both the contribution to the OPLS model, that is, concentration variant and correlation to the OPLS model, that is, concentration invariant. The current strategy focuses on this complex problem. This strategy also emphasizes additional information obtained when proper multivariate modeling is combined with proper and efficient visualization. OPLS-DA models using S-plots will be contrasted with the use of a Student's *t*-test[16] for single metabolite identification. The similarities with the useful STOCSY[17,18] plot will be discussed. An additional complication discussed is the multiclass situation, exemplified here by two transgenic and one wild-type (WT) lines. Building a model based on all three classes simultaneously will result in a reference point, which is a mixture of all class data used in the model. Since all plots and graphs will be anchored to this artificial point, this causes problems for visualization and biological interpretation of the model. The use of the SUS-plot that combines information from a number of two-class models having the same reference is suggested as a solution.

The OPLS-DA[19] approach as all other regression methods is sensitive to model complexity. Therefore, cross-validation (CV) was used to estimate the relevant number of components in the OPLS models.

For demonstrating the new multivariate strategy, two different transgenic poplar lines were compared to wild-type poplar by gas chromatography−mass spectrometry (GC/MS) metabolomics. The two transgenic lines, 2B and 5, have been up- and downregulated for the expression of the *PttPME1* gene, respectively. This regulation was expected to affect the degree of methyl esterification (DM) of homogalacturonan, the most important component of pectin in plant cell walls. Full biological information about these lines is published elsewhere.[20] It has been shown that line 5 was more affected in DM and only a small difference could be seen between line 2B and WT. Nevertheless, the metabolic profile was of interest as both lines indicated several symptoms of oxidative stress responses, which are typically associated with both biotic and abiotic stress.[21]

EXPERIMENTAL SECTION

Sample Collection and Preparation. Transgenic poplar lines 5 and 2B of *Populus tremula* L. *x tremuloides Michx* were obtained and grown along with the WT poplar clone T89 in the greenhouses of the Umeå Plant Science Center.[20] When plants were ~2 m tall, the internode 42 (counting from the top) was isolated, frozen in liquid nitrogen, and debarked after partial thawing, and developing xylem tissue was scraped into a liquid nitrogen vessel. A number of trees were obtained for each line, 7 plants for line 5, 9 plants for line 2B, and 10 plants for WT. All samples were ball-milled to a fine powder in a frozen condition and then lyophilized. The xylem powder (5.0 mg) was extracted in a vibration mill with 1 mL of extraction mix consisting of water (20%), methanol (60%), chloroform (20%), and 11 internal standards (7.5 ng/μL).[22] The samples were then derivatized by shaking them with 30 μL of methoxyamine hydrochloride (15 mg mL^{-1}) in pyridine for 10 min at 5 °C and then incubating them for 16 h at room temperature. The samples were then trimethylsilylated by adding 30 μL of *N*-methyl-*N*-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane and incubating them for 1 h at room temperature. After silylation, 30 μL of heptane was added.

GC/MS Analysis. The samples were analyzed according to Gullberg et al. by GC/time-of- flight (TOF)-MS together with blank control samples and a series of *n*-alkanes (C12−C40) to allow retention indexes to be calculated.[23] A 1-μL aliquot of each derivatized sample was injected splitless, by an Agilent 7683 autosampler (Agilent, Atlanta, GA) into an Agilent 6890 gas chromatograph equipped with a 10 m × 0.18 mm i.d. fused-silica capillary column with a chemically bonded 0.18-μm DB 5-MS stationary phase (J&W Scientific, Folsom, CA). The injector temperature was 270 °C, the septum purge flow rate was 20 mL min^{-1}, and the purge was turned on after 60 s. The gas flow rate through the column was 1 mL min^{-1}; the column temperature was held at 70 °C for 2 min, then increased by 40 °C min^{-1} to 320 °C, and held there for 2 min. The column effluent was introduced into the ion source of a Pegasus III time-of-flight mass spectrometer, GC/TOF/MS (Leco Corp., St. Joseph, MI). The transfer line and the ion source temperatures were 250 and 200 °C, respectively. Ions were generated by a 70-eV electron beam at an ionization current of 2.0 mA, and 30 spectra s^{-1} were

[15] Kvalheim, O.; Karstang, T. *Chemom. Intell. Lab. Syst.* **1989**, *2*, 37−52.

[16] Student. *Biometrica* **1908**, *6* (1), 1−25.

[17] Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77* (5), 1282−1289.

[18] Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Anal. Chem.* **2005**, *77* (2), 517−526.

[19] Bylesjö, B.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2006**, *20* (8−10), 341−351.

[20] Siedlecka, A.; Wiklund, S.; Péronne, M. A.; Micheli, F.; Leśniewska, J.; Sethson, I.; Edlund, U.; Richard, L.; Sundberg, B.; Mellerowicz, E. J. Pectin methyl esteras inhibits intrustive and symplastic cell growth in developing wood of *populus* trees. Umeå Plant Science Center, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden. Manuscript in preparation.

[21] Öhman, D.; Siedlecka, A.; Lesniewska, J.; Wiklund, S.; Kleczkowski, L.; Sundberg, B.; Mellerowicz, E. J. Umeå Plant Science center, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden. Manuscript in preparation.

[22] Gullberg, J.; Jonsson, P.; Nordström, A.; Sjöström, M.; Moritz, T. *Anal. Biochem.* **2004**, *331* (2), 283−295.

[23] Schauer, N.; Steinhauser, D.; Strelkov, S.; Schomburg, D.; Allison, G.; Moritz, T.; Lundgren, K.; Roessner-Tunali, U.; Forbes, M. G.; Willmitzer, L.; Fernie, A. R.; Kopka, J. *FEBS Lett.* **2005**, *579*, 1332−1337.

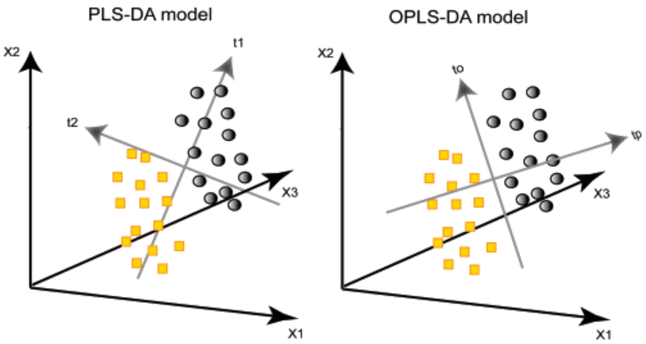

Figure 1. Geometrical illustration of the difference between PLS-DA and OPLS-DA. The OPLS-DA model is rotated so that between class variation, difference between squares and circles, is found in the predictive component t_p and within class variation is seen in the first y-orthogonal components t_o.

recorded in the mass range 50−800 m/z. The acceleration voltage was turned on after a solvent delay of 170 s. The detector voltage was 1660 V.

Deconvolution and Preprocessing. All data were processed using the hierarchical multivariate curve resolution (H-MCR) MATLAB script.[24] The H-MCR deconvoluted GC/MS data allow the mass spectra of all detected compounds to be compared with spectra in National Institute of Standards and Technology (NIST) library and in-house database. In addition, the peak area from the pure profiles was calculated by the H-MCR scripts resulting in 81 metabolites, of which 55 were identified. The resulting data matrix was normalized using the concentrations of 11 added internal standards.[25] After normalization, the standards were removed so that the data used for modeling consisted of extracted compounds.

OPLS Modeling. Partial least-squares to latent structures, PLS, is a regression method commonly used in multivariate studies in order to find the relationship between two data tables referred to as X, here the GC/MS data, and Y, here a binary vector with the value 0 for WT class and 1 for the transgenic class (line 2B or line 5). OPLS, is a modification of PLS, which separates the systematic variation in X into two parts, one that is linearly related to Y and one that is orthogonal to Y. A geometrical illustration of the differences is demonstrated (Figure 1).[12,14] Hence, the OPLS model comprises two blocks of modeled variation: (1) the Y-predictive ($\mathbf{T}_p\mathbf{P}_p{}^T$) block, which represents the *between class* variation, and (2) the Y-orthogonal ($\mathbf{T}_o\mathbf{P}_o{}^T$) block also referred to as the uncorrelated variation, which constitutes the *within class* variation. In the formulas above, **T** represents the score matrix and **P** the loading matrix. Figure 1 is further visualizing the advantage obtained by the optimal rotation by OPLS as compared to PLS.

The OPLS modeling, with the above problem formulation of a WT and two transgenic lines, will result in two models one comparing line 5 with WT and the second comparing line 2B with WT. All models were calculated using SIMCA-P+11.0, Umetrics.

Transformation and Scaling. Noise and artifacts are common in metabolomics data, which makes multivariate projection models sensitive to the choice of scaling and transformation. In this example, column centering and pareto scaling ($1/\sqrt{\text{SD}}$, where SD is the standard deviation) for each variable within each model was used prior to modeling.[26] Pareto scaling is a technique that is a golden mean of column centering and scaling to unit variance (centering + 1/SD), UV. The advantage of using pareto scaling compared to UV is that it reduces the impact of noise and artifacts in the models, which is positive for the models' predictive ability.

Model Dimensionality and Cross-Validation. The model dimensionality (number of significant OPLS components) is important both for the interpretation and for the predictive ability of the model. In the present study, CV was used according to the following strategy. During CV, one sample from the smaller class and one or two from the larger class, selected in randomized order, were excluded for each CV round to obtain a balanced class size. Thus, each sample was excluded once and only once. There is an extensive literature on CV and validation of multivariate models that demonstrates that different CV strategies can give slightly different results, but this is beyond the scope of this paper.[27,28]

Visualization. The use of appropriate visualization tools helps communication and interpretation of scientific data.[29] We have used five types of plots to improve the interpretation by means of visualization. The following plots were presented: cross-validated score plots, S-plot, loading plots with confidence intervals, SUS-plot, and raw data plots. In cross-validated score plots, each sample, i, was represented by two points, one for the cross-validated score value ($t_{i,cv}$) and one for the model score value ($t_{i,p}$). This procedure will make CV more transparent.

The S-plot visualizes the variable influence in a model. It is a scatter plot that combines the covariance and correlation loading profiles resulting from a projection-based model, e.g., the predictive component, t_p, of an OPLS-DA model. This corresponds to combining the contribution or magnitude (covariance) with the effect and reliability (correlation) for the model variables with respect to model component scores. This combination is not unique for OPLS but can be applied to other projection-based models, e.g., PCA or PLS. The combination of correlation and covariance in one plot has also been successfully applied with 1-D NMR data, the STOCSY plot, to improve interpretation of the predictive component.[17,18,30] The two vectors used in the S-plot were calculated as

$$\text{Cov}(t,X_i) = \frac{t^T X_i}{N-1} \tag{1}$$

$$\text{Corr}(t,X_i) = \frac{\text{Cov}(t,X_i)}{s_t s_{X_i}} \tag{2}$$

where t is the score vector in the OPLS-DA model, i is the centered variable in data matrix **X**, and s defines the estimated

[24] Jonsson, P.; Johansson, E. S.; Wuolikainen, A.; Lindberg, J.; Schuppe-Koistinen, I.; Kusano, M.; Sjöström, M.; Trygg, J.; Moritz, T.; Antti, H. *J. Proteome Res.* **2006**, *5* (6), 1407−1414.

[25] Jiye, A.; Trygg, J.; Gullberg, J.; Johansson, A. I.; Jonsson, P.; Antti, H.; Marklund, S. L.; Moritz, T. *Anal. Chem.* **2005**, *77* (24), 8086−8094.

[26] Wold, S.; Johansson, E.; Sjöström, M.; Cocchi, M. *PLS*, *In;* Escom Science: Leiden, 1993; pp 523−550.

[27] Wold, S. *Technometrics* **1978**, *20*, 397−405.

[28] Xu, Q. S.; Liang, Y. Z. *Chemom. Intell. Lab. Syst.* **2001**, *56* (1), 1−11.

[29] Cleveland, W. S. *The elements of graphing data*; Wadsworth Publ. Co.: Belmont, CA, 1985.

[30] Holmes, E.; Loo, R. L.; Cloarec, O.; Coen, M.; Tang, H. R.; Maibaum, E.; Bruce, S. J.; Chan, Q.; Elliott, P.; Stamler, J.; Wilson, I. D.; Lindon, J. C.; Nicholson, J. K.*Anal. Chem.* **2007**, *79* (7), 2629−2640.

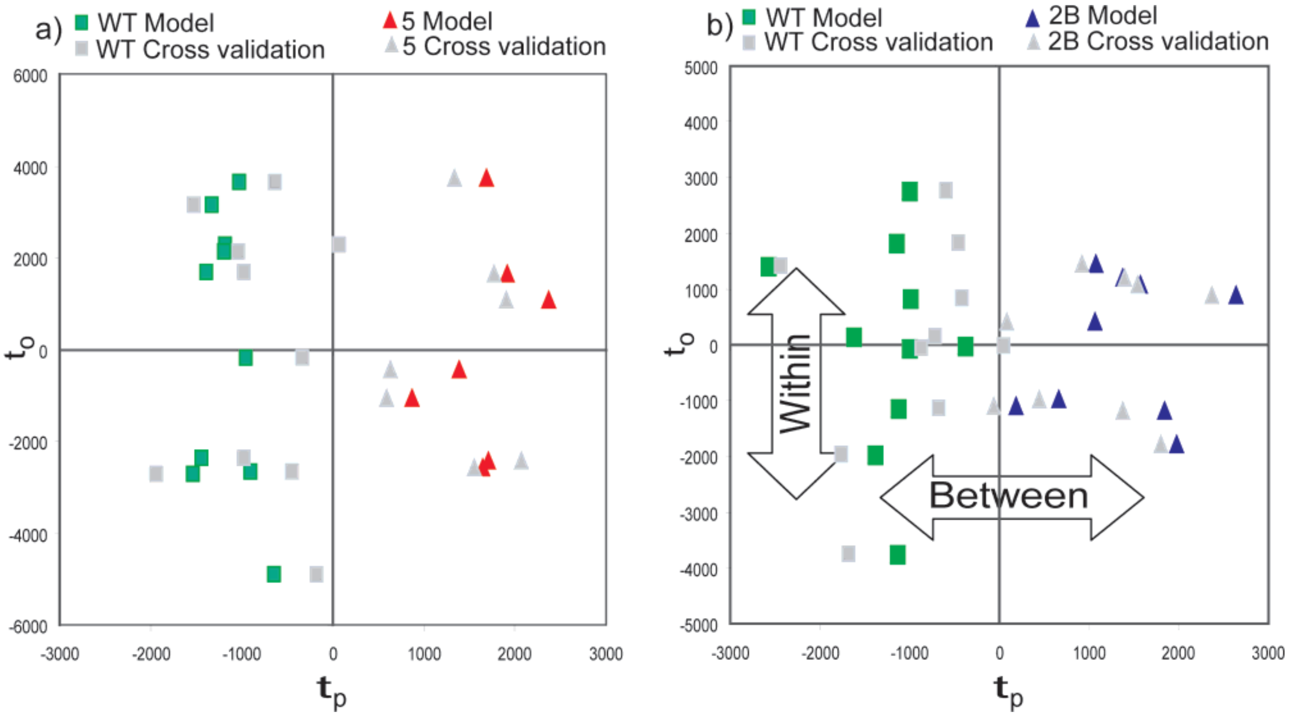

Figure 2. Cross-validated score plots for OPLS-DA models (a) line 5 vs WT and (b) line 2B vs WT. As the plot focus on the modeled score, $t_{i,p}$, and CV score, $t_{i,cv}$, observation i have identical orthogonal score, t_o.

standard deviation. Hence, Cov(t,X) and Corr(t,X) are vectors of the same length as number the of variables in the model. These two vectors are plotted in a scatter plot and are S-shaped unless the variable variance is uniform. For implementation of the two formulas 1 and 2 with the NIPALS algorithm, use the one described in Supporting Information. The x-axis, Cov(t,X), in the S-plot is a visualization of contribution (covariance) and the y-axis, Corr(t,X), spans between ±1 as the correlation (reliability) has a theoretical minimum of −1 and a maximum of +1. The statistical S-plot is used for identification of possible biochemically interesting compounds for the predictive variation as well as the orthogonal variation. A complimentary tool for identification of interesting compounds is to plot the loading vector, Cov(t_p,X), with its corresponding jack-knifed confidence intervals as these provide additional information about metabolite variability.[31] With GC/MS or LC/MS data, the loading plot is preferably sorted by size in order to separate up- and downregulated metabolites in each end of the plot.

To compare the outcome of two or more OPLS models, the SUS-plot plays an important role in visualizing both the shared and unique information. The SUS-plot combines the Corr(t_p,X) profiles from two models, in one 2-D plot. As the SUS-plot displays correlation, it should be scaled between −1 and +1 for both axes. By expanding to 3-D plots or multiple 2-D plots, an additional number of models can be compared and visualized simultaneously. Metabolites close to the diagonal will be shared between classes and metabolites outside the diagonal will be unique for the specified class; see results in Figure 6.

RESULTS

Model Quality. OPLS-DA was carried out between lines 5 and WT samples in one model and between lines 2B and WT samples in a separate model. The first model (5 vs WT) resulted in one predictive and three orthogonal (1 + 3) components with the cross-validated predictive ability $Q^2(Y) = 77\%$, the total explained variance $R^2(X) = 73\%$, and the variance related to class separation $R^2_p(X) = 11\%$. The second model (2B vs WT) resulted in one predictive and one orthogonal component (1 + 1). The predictive ability was $Q^2(Y) = 64\%$, the total explained variance $R^2(X) = 42\%$, and the variance related to the differences between the two classes $R^2_p(X) = 13\%$.

Cross-Validated Score Plots. The between class variation, t_p, and the CV predictive ability, $Q^2(Y)$, were directly visualized in the CV score plots. The model containing line 5 and WT had a slightly higher $Q^2(Y)$ and the separation between the classes appeared acceptable as indicated by only one sample being misclassified, (Figure 2a). In the model containing 2B and WT, a few samples were on the borderline between the two classes (Figure 2b). This misclassification indicates a higher uncertainty considering class separation compared to differentiation between line 5 and WT. The resulting CV estimate, i.e., $Q^2(Y)$, was lower for this model. The first uncorrelated component, t_o, in the model containing line 5 and WT indicated two subgroups within the wild type, (Figure 2a). The interpretation of between class variation, i.e., class separation, $t_{p,}$ and within class variation, t_o, is facilitated by the S-plot. Deviating samples with a large difference between the model score value ($t_{i,p}$) and the cross-validated score value ($t_{i,cv}$) need additional analysis as the distance indicates the uncertainty of the prediction.

[31] Efron, B.; Gong, G. *Am. Statistician* **1983**, *37*, 36−48.

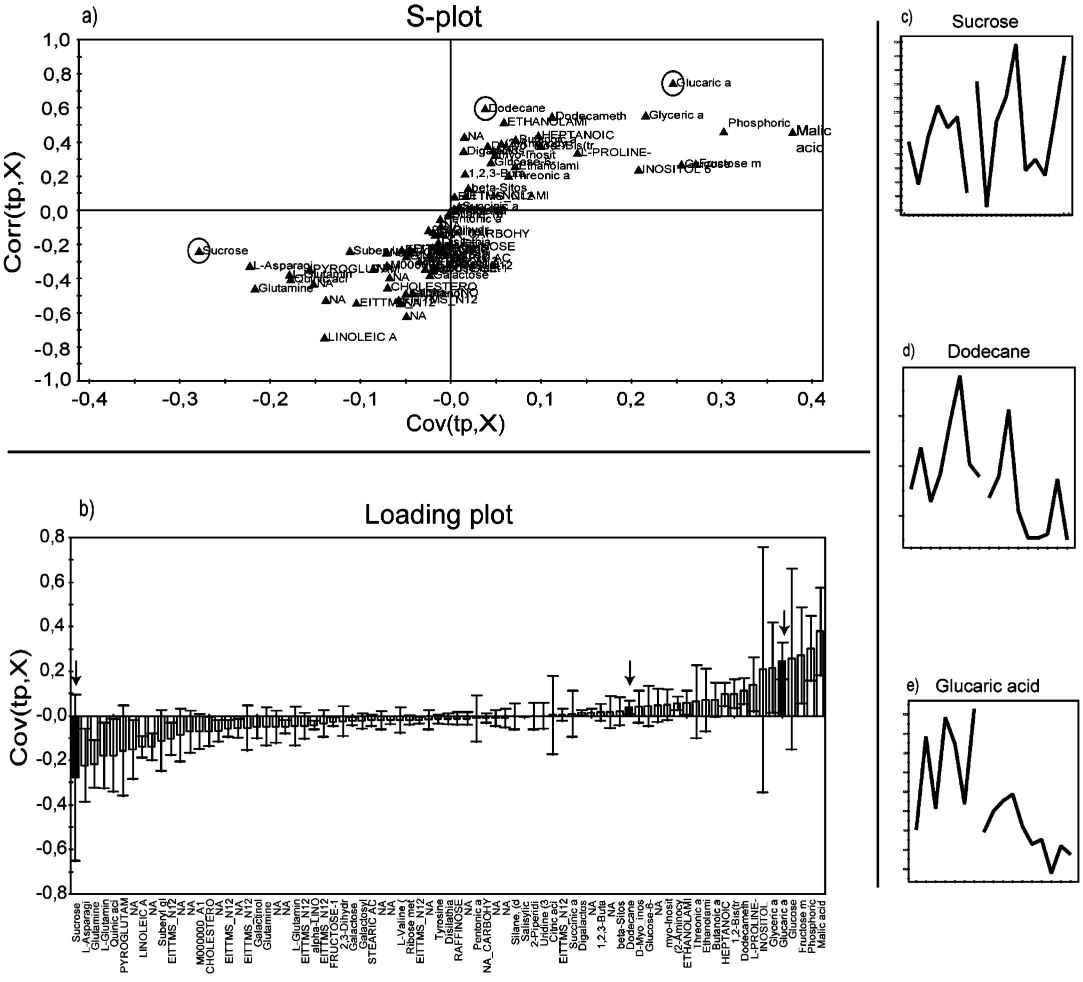

Figure 3. Strategy for identification of interesting metabolites for line 5. (a) S-Plot, three metabolites are highlighted by circles to demonstrate different regions in the S-plot. These are shown in (c−e). (b) Loading plot with jack-knifed confidence intervals. The arrows indicate the metabolites shown in (c−e). (c) Raw data for sucrose located in a high-risk region. (d) Raw data for dodecane located in a uncertain and difficult region. (e) Raw data for glucaric acid located in a low-risk region. Of these three highlighted metabolites, the only one selected as an interesting biomarker was glucaric acid.

S-Plot for Predictive Component, Class Separation. Statistically significant metabolites related to the differences between line 5 and WT were selected from the S-plot (Figure 3a) together with the jack-knifed confidence interval (Figure 3b). The jack-knifed confidence interval, $\mathrm{CIJF_{JK}}$, is calculated as

$$\mathrm{CIJF_{JK}} = \mathrm{SEcv}\ t(\alpha,\mathrm{df})$$

where $\mathrm{SE_{CV}}$ is the standard error for each loading based on CV, t is the statistical value based on $\alpha = 0.05$, and df is the degrees of freedom based on the number of CV rounds. Focusing the selection of interesting metabolites solely on the covariance will lead to a selection biased toward metabolites with a high spectral intensity, which often is related to a high concentration.[25] This might be one reason that the term “the usual suspects” has been coined in metabolomics.[32] Metabolomic data with a large dynamic range in concentration will usually have few metabolites with a large concentration but many with a low or very low concentration. To base the selection of interesting metabolites solely on the correlation will result in a selection where a number of the biochemical compounds will have very low concentration. The risk for false positives (type I error) thereby increases. The selection of potentially biochemically interesting compounds therefore needs a combination of covariance and correlation information, which is the purpose of the S-plot. For the separation between line 5 and WT, three metabolites were highlighted in the S-plot, sucrose, glucaric acid, and dodecane. The plot shows that sucrose had a low correlation, which means a low reliability, for class separation. The same low reliability was also found in the loading plot where the jack-knifed based confidence interval includes zero. The raw data plot confirmed this with both classes overlapping. In contrast, the glucaric acid had a high correlation in the S-plot and was therefore reliable for class separation. Similar information was found in the loading plot where it had a small confidence interval. In the S-plot, dodecane was located in a region with ambiguous significance and it was not selected as it was not supported by the confidence interval.

[32] Robertson, D. G. *Toxicol. Sci.* **2005**, *85*, 809−822.

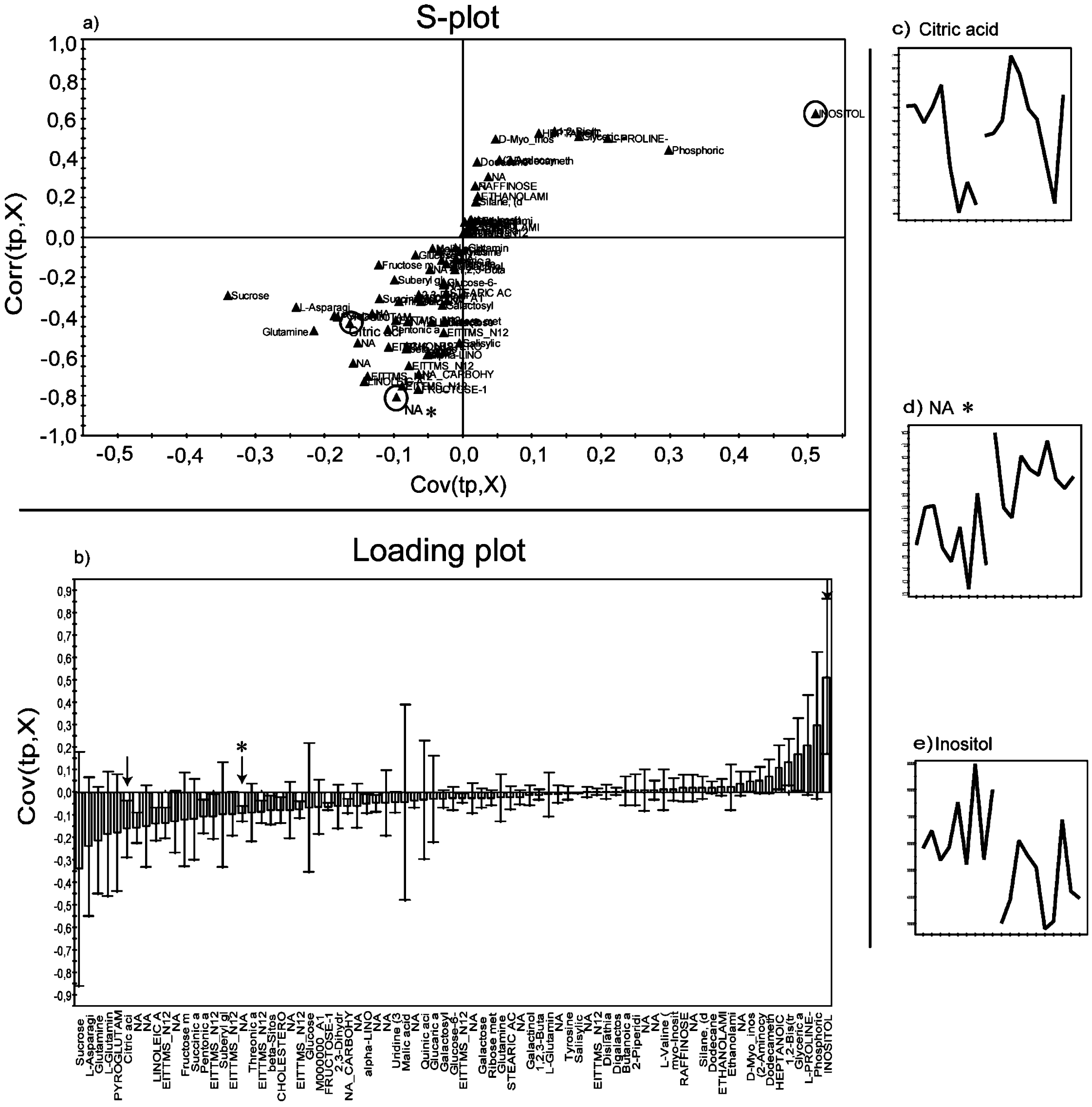

Figure 4. Strategy for identification of interesting metabolites for line 2B. (a) S-Plot, three metabolites are highlighted by circles to demonstrate different regions in the S-plot. These are shown in (c−e). (b) Loading plot with jack-knifed confidence intervals. The arrows indicate the metabolites shown in (c−e). (c) Raw data for citric acid, which is located in an uncertain region. (d) Raw data for a not assigned metabolite (NA*) located in a region that demands check of the raw data since it is a low-risk region in the Corr(t_p,X) but uncertain in the Cov(t_p,X). (e) Raw data for inositol located in a low-risk region. Of these three highlighted metabolites, inositol and the NA* metabolites were selected as interesting biomarkers.

Interesting compounds for the line 2B were selected from the S-plot combined with the jack-knifed confidence interval (Figure 4a,b). In the S-plot, three metabolites are highlighted, citric acid, one not assigned (NA), and inositol, (Figure 4c−e). One of these metabolites, citric acid, was located in another ambiguous region of the S-plot. This metabolite was not selected since the confidence interval in the loading plot did not support this selection. The other two metabolites were in a low-risk region, and they were selected for further investigation of their biochemical significance.

S-Plot of Orthogonal Component. The orthogonal component from the model line 5 versus WT was interpreted as a possible subgrouping in the WT class. Compounds related to the variation within the WT class were detected in the S-plot from the orthogonal component (Figure 5). Sucrose, with both high correlation and high covariance, was the metabolite creating most of this subgrouping within the WT class. In order to identify further metabolites for the subgrouping, one should use the same strategy of comparing both covariance and correlation as for the predictive component.

SUS-Shared and Unique Structure. When comparing two or more models, the interest of finding shared as well as unique compounds between a set of different models is of great importance for the total understanding of a drug action or a gene modification. In the SUS-plot, the correlation from the predictive component, Corr(t_p,X), of each model was plotted against each other. To improve clarity, all points inside a threshold for each axis can be eliminated. In this example, all metabolites that were found not significant in either class according to the confidence interval seen in the loading plot, Figures 3 and 4, were removed. The unique effects will only be significant in either line 5 or 2B. Unique effects were found close to either the X or Y axis for line 5 and line 2b, respectively, while the shared effects were located on the diagonals (Figure 6). The unique identified metabolites were, for example, malic acid (+) and glucaric acid (+) for line

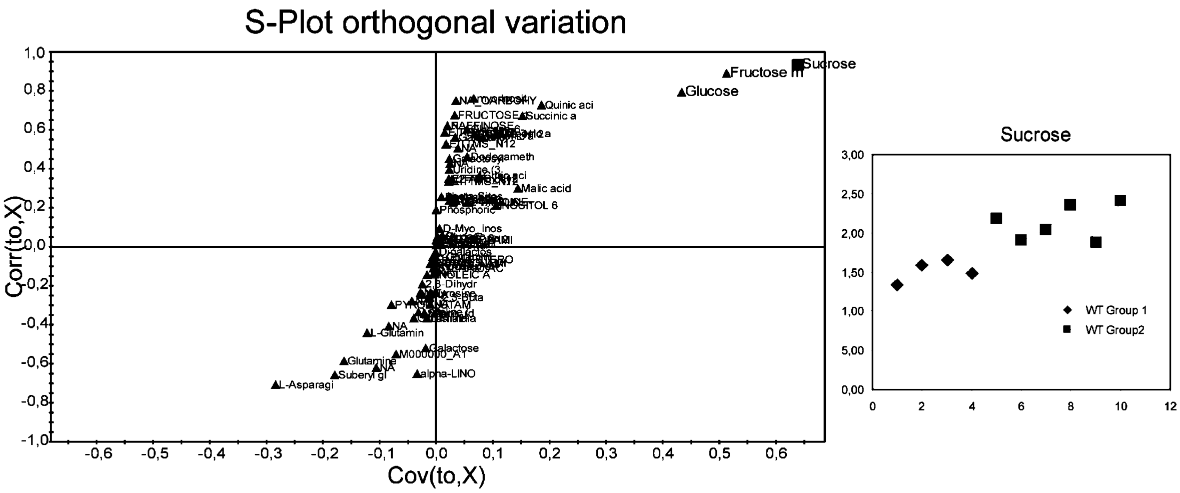

Figure 5. In the model line 5 vs WT. The class separation within the WT was interpreted by using the S-plot from the orthogonal component. One of the metabolites responsible for this separation was sucrose.

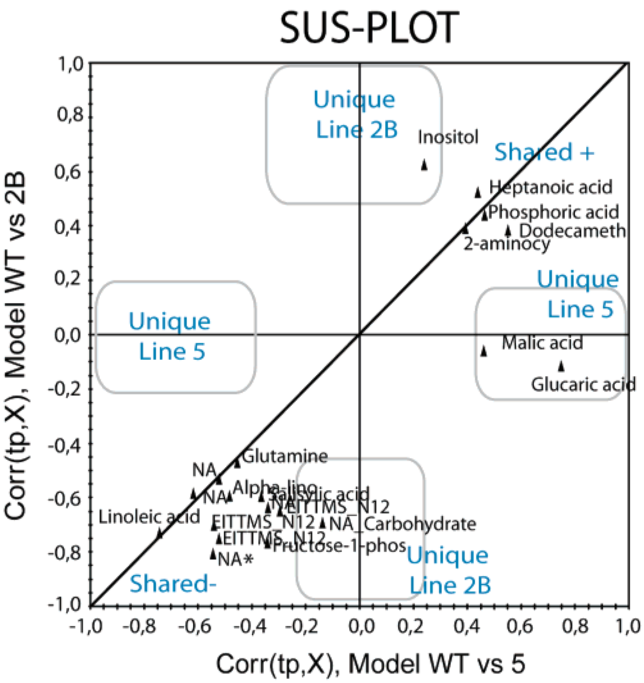

Figure 6. SUS-plot between line 5 and line 2B. Metabolites close to the diagonal line are equally affected in both lines. All metabolites that were found not significant according to the confidence interval seen in the loading plots, Figures 3 and 4 were removed in this plot to enhance visualization.

5, and inositol (+) for line 2B. Several metabolites were shared, for example, salicylic acid, which was downregulated in both lines.

DISCUSSION

Multivariate models have the advantage that they find relations among correlated variables, as is often the case in metabolomics studies, and models also have the ability to separate systematic variation from noise. OPLS has the additional advantage that it separates the predictive variation from the orthogonal variation and can be studied and interpreted separately. To get the full potential out of multivariate models, they must be combined with appropriate diagnostics and validation as well as an efficient visualization.

Practical Use of the S-Plot. The preferred selection of metabolites has a high covariance combined with a high correlation resulting in a small confidence interval. The higher the correlation the more reliable is the selection. The lower the covariance is, the larger the risk that the observed effect stems from analytical variation and noise. To give exact limits for the high-risk regions is however a dangerous and difficult task. This is because both correlation and covariation depend on the quality and stability of the data. When including more samples in the model, smaller correlations can be proved to be statistically true since the data stabilize. Thus, instead of using predefined limits, the quality of interpretation about interesting compounds depends on the experimental design, number of samples in the study, and what level of correlations that is of interest for further ivestigations.[33−35] OPLS-DA together with the S-plot and SUS-plot allows for mining from complex data to propose which metabolites are statistically and potentially biochemically interesting compounds.

The physiological interpretation of interesting compounds found in the predictive component, e.g., salicylic acid, inositol, andfructose 1-phosphate for line 2B and malic acid and glucaric acid for line 5 will be discussed in an additional paper. A tentative explanation about the shared metabolites, e.g., salicylic acid (important compound in both models), was that it provided important information regarding an unexpected common stress reaction of the plants of line 5 and line 2B, in spite of having opposite transgenic modification and exhibiting several opposite changes in growth and in cell wall composition.[21]

The findings in the orthogonal component are often related to experimental variation.[36] Here it was hypothesized that it was caused by the experimental scraping procedure.

Comparison of S-Plot with a *t*-Test Approach. The *t*-test is today widely used for selection of interesting compounds. This

[33] Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nyström, A.; Pettersen, J.; Bergman, R. *Chemom. Intell. Lab. Syst.* **1998**, *42* (1−2), 3−40.

[34] Cohen, J. *Am. Phychologist* **1990**, *45* (12), 1304−1312.

[35] Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for experimenters An introduction to design, data analysis and model building*; John Wiley & Sons: New York, 1978.

[36] Bylesjö, M.; Eriksson, D.; Sjödin, A.; Jansson, S.; Moritz, T.; Trygg, J. *BMC Bioinformatics*. In press.

strategy appears inappropriate for several reasons. The major objection is that the *t*-test gives no consideration to variable intensity, which often is related to metabolite concentration.[25] For much the same reason, in the present study we have chosen pareto scaling as it reduces the impact of noise and artifacts in the models. Additionally, no clue regarding the confidence interval or orthogonal variation (e.g., sucrose in WT) will be discovered by the *t*-test.

The American statistician and psychologist Jacob Cohen stated in his publication in1994, The Earth Is Round ($p < 0.05$), that "Everyone knows that confidence intervals contain all the information to be found in significance test and much more. They not only reveal the status of the trivial nil hypotheses but also the status of non-nil hypothesis and thus remind researchers about the possible operation of the crude factor. Yet they are rarely to be found in the literature".[37] Multivariate models such as PLS and OPLS include both statistical significance based on cross-validation and confidence intervals based on jack-knifing estimations as well as magnitude and reliability of the data provided by good visualization.

Additional Use of OPLS-DA and S-Plot. We have demonstrated the use for GC/MS metabolomics data. The methodology itself is independent of the analytical platform and can be extended to LC/MS, 1-D NMR, 2-D NMR, etc. Extension into analysis of proteomics, transciptomics, and genomics appears straightforward. The general approach to handle problems with three or more classes by combining the results from two models can be extended to dynamic data, i.e., data collected at different time points as well as data from formalized design of experiments.

CONCLUSIONS

In the present study, we have designated an OPLS model-based approach for definition of statistically and potentially biochemically significant compounds found in the up- and downregulated *PttPME1* modified poplar plants. The approach includes the description of two key plots named S- and SUS-plots. These two plots have the potential to enhance interpretation of complex data structures. The used method appears parsimonious and robust, with general applicability for data mining from metabolomic and similar data.

ACKNOWLEDGMENT

This work was supported by grants from Swedish Research Council, Swedish Foundation for Strategic Research, MKS Umetrics, and the EU-Strategic Funding. The authors thank Krister Lundgren and Inga-Britt Carlsson for technical assistance and Anna Wuolikainen for help with the GC/MS processing using the H-MCR MATLAB script.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review June 26, 2007. Accepted October 1, 2007.

AC0713510

[37] Cohen, J. *Am. Psychologist* **1994**, *49* (12), 997–1003.