Metabolomics:
A Tool for Studying Plant Biology

Jonas Gullberg
Faculty of Forest Science
Department of Forest Genetics and Plant Physiology
Umeå

Doctoral thesis
Swedish University of Agricultural Sciences
Umeå 2005

Acta Universitatis Agriculturae Sueciae
2005: 88

ISSN 1652-6880
ISBN 91-576-6987-2

Tryck: Arkitektkopia, Umeå, Sweden, 2005

Abstract

Gullberg, J. 2005. *Metabolomics: A Tool for Studying Plant Biology*.
Doctor's dissertation.
ISSN 1652-6880, ISBN 91-576-6987-2

In recent years new technologies have allowed gene expression, protein and metabolite profiles in different tissues and developmental stages to be monitored. This is an emerging field in plant science and is applied to diverse plant systems in order to elucidate the regulation of growth and development. The goal in plant metabolomics is to analyze, identify and quantify all low molecular weight molecules of plant organisms. The plant metabolites are extracted and analyzed using various sensitive analytical techniques, usually mass spectrometry (MS) in combination with chromatography. In order to compare the metabolome of different plants in a high through-put manner, a number of biological, analytical and data processing steps have to be performed. In the work underlying this thesis we developed a fast and robust method for routine analysis of plant metabolite patterns using Gas Chromatography-Mass Spectrometry (GC/MS). The method was performed according to Design of Experiment (DOE) to investigate factors affecting the extraction and derivatization of the metabolites from leaves of the plant *Arabidopsis thaliana*. The outcome of metabolic analysis by GC/MS is a complex mixture of approximately 400 overlapping peaks. Resolving (deconvoluting) overlapping peaks is time-consuming, difficult to automate and additional processing is needed in order to compare samples. To avoid deconvolution being a major bottleneck in high through-put analyses we developed a new semi-automated strategy using hierarchical methods for processing GC/MS data that can be applied to all samples simultaneously. The two methods include base-line correction of the non-processed MS-data files, alignment, time-window determinations, Alternating Regression and multivariate analysis in order to detect metabolites that differ in relative concentrations between samples. The developed methodology was applied to study the effects of the plant hormone GA on the metabolome, with specific emphasis on auxin levels in *Arabidopsis thaliana* mutants defective in GA biosynthesis and signalling. A large series of plant samples was analysed and the resulting data were processed in less than one week with minimal labour; similar to the time required for the GC/MS analyses of the samples.

Keywords: Plant metabolomics *Arabidopsis thaliana*, Chromatography mass-spectrometry, Design of experiment, Deconvolution, Chemometric analysis, Gibberellin

Author's address: Jonas Gullberg, Umeå Plant Science Centre (UPSC), Department of Forest Genetics and Plant Physiology, SLU, S-901 83, UMEÅ, Sweden

Contents

Appendix

List of Papers

This thesis is based on the following papers, which are referred to in the text by the corresponding Roman numerals:

I. Gullberg, J., Jonsson, P., Nordström, A., Sjöström, M. & Moritz, T. 2004. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. *Analytical Biochemistry 331*, 283-295.

II. Jonsson, P.*, Gullberg, J.*, Nordström, A., Kusano, M., Kowalczyk, M., Sjöström, M. & Moritz, T. 2004. A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Analytical Chemistry 76*, 1738-1745.

III. Jonsson, P., Johansson, AI., Gullberg, J., Trygg, J., A, J., Grung, B., Marklund, S., Sjöström, M., Antti, H. & Moritz, T. 2005. High through-put data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyzes. *Analytical Chemistry,* In Press.

IV. Gullberg, J., Wiklund, S., Trygg, J., Sjöström, M., Harberd, NP. & Moritz, T. Effect of gibberellin signalling on IAA levels and metabolite profiles in *Arabidopsis*. (Manuscript)

*To be considered joint first authors

Introduction

Plants must be able to adapt to survive changes in the external conditions, such as fluctuations in temperature, light, drought, nutrient supplies and attacks by pests during their growth and development. Diverse environmental signals are integrated in plant developmental programs in adaptive responses that maximize their competitiveness and survival. A wide range of plant systems have been used to study plant functions. *Arabidopsis thaliana* is the most intensively studied plant model system in functional biology today and its use has dramatically increased our knowledge of developmental processes in plants (Meinke *et al.*, 1998). *Arabidopsis* is a small annual plant with a rapid life cycle (germination to mature seed in six weeks), it has a small and sequenced genome, and a large collection of mutants is available (The Arabidopsis Genome Initiative, 2000).

In order to understand the regulation of growth and development of plants such as *Arabidopsis*, the expression of genes, proteins and metabolites are studied in different tissues and development stages. Global analysis of genes, proteins and metabolites [transcriptomics (Andersson *et al.*, 2004), proteomics (Newton *et al.*, 2004) and metabolomics, respectively (Fiehn, 2002; Sumner, Mendes & Dixon, 2003)] are emerging fields in plant science. The genes, transcripts and proteins are regulated and organized in a complex network that controls plant development. To be able to apply these approaches to plant biology on a routine basis, appropriate methodology has to be developed. Transcriptomic analysis is today quite straightforward, but both proteomics and metabolomics are developing fields. For example, metabolomic analysis must be performed using highly sensitive analytical instruments (e.g. mass spectrometry, MS, in combination with chromatography) to give interpretable results.

A further key to understanding the growth and development of plants is knowledge of plant hormones, e.g. auxins, cytokinins, gibberellins (GAs), abscisic acid and ethylene, and their metabolism (Davies, 2004). The standard definition of a hormone is an organic compound produced by one tissue in an organism and transported to another tissue, where it induces a specific physiological response (Lawrence, 1995). However, in the context of plant hormones they can also work as plant growth regulators at the site of their action. Cross-talk between different growth-regulating substances increases the complexity and levels of control of growth and development.

Primary metabolites and hormones are affected by, and affect, the physiological processes in the developmental processes in the plant and can be seen as effect of end products of gene expression.

In this thesis I consider the development of extraction protocols, data comparison and analytical methods for high-throughput global metabolite screens. The developed methodology has been applied to study the effects of the plant hormone GA on the metabolome, with specific emphasis on auxin levels in *Arabidopsis thaliana* mutants defective in GA biosynthesis and signalling.

Background

Metabolomics

The goal in plant metabolomics is to comprehensively analyze, identify and quantify the metabolome of plants (Fiehn, 2002). The metabolome is defined as all low molecular weight molecules (metabolites) present in a cell (Fiehn, 2002; Harrigan & Goodacre, 2003; Sumner, Mendes & Dixon, 2003). The metabolites can be viewed as the end products of gene expression and enzymatic activity. Thus**,** metabolomics has been proposed as a useful tool for studying gene function. In recent years metabolomics has become a complementary method to large-scale analysis of gene transcript levels (microarray analysis, transcriptomics) and proteins (proteomics). The general aim of these large-scale analyses is to obtain information that can explain and identify the differences between certain sets of organisms (e.g. differences in genotypes), or to elucidate factors that influence biochemical events. The assumption in functional biology is that a change in the transcriptome (the complete collection of transcribed elements of the genome) affects the catalytic activities of enzymes, causing a change in the metabolome. The size of the metabolome differs between different organisms, and the plant kingdom has been estimated to produce up to 200,000 metabolites in total (Fiehn, 2002). However, specific plant and tissues contain fewer metabolites. *Arabidopsis* leaves have been estimated to contain approximately 5000 different primary and secondary metabolites (Bino *et al.*, 2004). The plant metabolome contains organic species – such as amino acids, fatty acids, carbohydrates, organic acids and lipids (see also KEGG, for a partial classification of the compounds http://www.genome.ad.jp/kegg/catalog/compounds.html, 19-August-2005), both elemental (Lahner *et al.*, 2003) and inorganic species. The diversity of the metabolome is much more complex in comparison to the 20 different amino acids present in a protein and the four nucleotide bases in the DNA sequence. Furthermore, the differences in the concentration of components in the metabolome is estimated to vary from pmol–mmol (Dunn & Ellis, 2005). Therefore, it is impossible to analyze all metabolites in a single analysis with current analytical equipment. To address this problem different analytical approaches have been developed to answer specific types of biological questions (Fiehn, 2002). As described above, metabolomics is an unbiased method that has to be selective and sensitive. This is in contrast to the more traditional way to analyze metabolites, e.g. analysis of targets for plant hormones (Ljung *et al.*, 2005). The qualitative and quantitative analysis in target analysis is focused on only a few metabolites. For each metabolite a corresponding internal standard is often used and the purified extract is separated and analyzed by a sensitive detection method. Metabolite profiling or metabolic profiling is focused on selected metabolites, for example a fraction containing a specific class of compounds (Broeckling *et al.*, 2005). Metabolic fingerprinting is a rapid and global analysis, which often does not involve chromatographic separation and metabolites are generally not identified (Choi *et al.*, 2004). Metabonomics is a similar strategy to metabolomics, but is often used in toxicology studies with NMR for global metabolite screenings (Nicholson *et al.*, 2002). The results of the different approaches give different levels of precision, depending on the number of metabolites analyzed. The earliest metabolite profiling focused on drug metabolites

(Horning & Horning, 1971). The first large-scale plant metabolite profiling was first performed by Roessner *et al.* (2000, 2001) on potato tubers (*Solanum tuberosum*) and *Arabidopsis thaliana* leaf extracts (Fiehn *et al.*, 2000a) using Gas Chromatography-Mass Spectrometry (GC/MS). Both metabolic profiling and metabolomics have been applied to many different plant species and to address different biological questions. To give some examples, the techniques have been used to study the metabolome of rice leaves (Sato *et al.*, 2004), changes in the metabolome during cold acclimation of *Arabidopsis* (Cook *et al.*, 2004), in comparisons of metabolic profiles of alfalfa and *M. truncatula* to identify new saponins (Huhman & Sumner, 2002), studies of growth processes in hybrid aspen (*Populus tremula* L. x *tremuloides* Michx.; Wiklund *et al.*, 2005) and characterization of flavonoid glycosides in genetically modified tomato (Le Gall *et al.*, 2003b). In addition, metabolomic approaches have been applied in flux analyses to elucidate plant metabolism (reviewed by Fernie, Geigenberger & Stitt, 2005), to determine stresses in genetically modified (GM) plants; Le Gall *et al.*, 2003a) and nutrition research (German, Roberts & Watkins, 2003). Further applications are reviewed by Sumner, Mendes & Dixon (2003) and Bino *et al.* (2004). In the field of human and animal metabolomics one of the main goals is to find metabolic biomarkers in tissues and biofluids that can act as disease indicators (Harrigan & Goodacre, 2003; Robertson, 2005). For a historical perspective of plant and animal metabolomics, see Harrigan & Goodacre (2003) and Sumner, Mendes & Dixon (2003). To understand biological behaviour in a holistic way the analysis should be as universal as possible. This requires the integration of biology, chemistry, mathematics, biostatistics and bioinformatics to convert information of diverse types into useful and interpretable results.

Metabolic Pathways

Metabolomics can be used to explain the biochemical function of annotated genes. It can also be used to define phenotypes and to bridge the genotype-to-phenotype gap (Fiehn, 2002). Furthermore, metabolites can also be related by their molecular structure and the fact that that they are built up by other metabolites. This can be visualized in maps or biochemical pathways describing the linkage between metabolite reactions. Known metabolite relationships have been used to compile publicly available reference biochemical reference databases (Mueller, Zhang & Rhee, 2003; Lange & Ghassemian, 2005). These reference biochemical databases contain not only information about biochemical pathways, cellular and molecular processes, but also information about the proteins that catalyse the reactions and the genes that code for them. An important point to remember is that the database information is limited since it does not cover species-specific pathways and the resulting diagrams do not cover all of the side reactions. However, databases can still provide a powerful visualization tool for the biological context of functional information. Examples of such databases include the Kyoto Encyclopedia of Genes and Genomes maps (KEGG; http://www.genome.ad.jp/kegg/; 19-August-2005; Kanehisa *et al.*, 2002) and the *Arabidopsis* Information Resource (TAIR, http://www.arabidopsis.org:1555/ARA/server.html, 19-August-2005).

Instruments used for Metabolomics

Numerous analytical techniques have been used in the field of plant metabolomics to monitor and explore metabolic differences between biological samples. I will here describe the most widely used methods for plant metabolite analysis: GC /MS and Liquid chromatography-mass spectrometry (LC/MS). Other important analytical techniques include liquid chromatography-photodiode array detection-mass spectrometry (LC/PDA/MS; Huhman & Sumner, 2002), capillary electrophoresis-mass spectrometry (CE/MS; Soga *et al.*, 2003; Sato *et al.*, 2004), Fourier-transform ion-cyclotron mass spectrometry (FT/MS; Tohge *et al.*, 2005) and Nuclear magnetic resonance (NMR; Ward *et al.*, 2003; Wiklund *et al.*, 2005), but these approaches will not be discussed in this thesis and are described in a number of reviews, for example Sumner, Mendes & Dixon (2003) and Dunn & Ellis (2005). The specificity, sensitivity and structural range of the different methods vary substantially.

Gas Chromatography-Mass Spectrometry (GC/MS)

Hyphenated chromatography-mass spectrometry approaches, such as high performance liquid chromatography (HPLC) and gas chromatography (GC) in combination with MS enables good metabolite identification and quantification compared to many other methods for screening metabolites. GC/MS (Watson, 1997; de Hoffmann & Stroobant, 2002) provides a robust system with excellent separation capacities and high thought-put possibilities, and is therefore the most commonly used analytical technique for routine analyses in the field of plant metabolomics. The separation of the analytes in gas chromatography is dependent on analyte interactions with the stationary face and the boiling point. Only compounds that are volatile can be separated on a GC column, so non-volatile metabolites must be derivatized prior to analysis by GC. A common way to solve this problem is to derivatize polar compounds containing functional groups such as –OH, –SH or –NH. For more details see the chapter *Derivatization for GC/MS*. After derivatization a portion of the sample is introduced into the inlet of the GC instrument. For volatile plant metabolites, methods such as headspace techniques (Verdonk *et al.*, 2003; Vikram, Prithiviraj & Kushalappa, 2004; Lui *et al.*, 2005) can be used to introduce metabolites to the GC column.

The inlet temperature is often higher than 250°C, at which many metabolites are evaporated. Two different injection methods are most widely used; splitless and split (Watson, 1997). In splitless injections the whole sample is introduced onto the high resolution capillary column, which is preferable to split injections (where only a portion is used) for trace analyses (Watson, 1997). The length of the capillary column varies between 10 to 60 metres. The polarity of the column can also be varied by changing the phenyl stationary phases, such as DB-5, DB-50 and CPSil-8. Aspects of different instrumental parameters effecting the GC analysis of metabolomes of human serum and yeast fermentation broths have been investigated by O'Hagan *et al.* (2005), who showed that optimization of GC conditions is required to improve analytical performance in metabolomic analysis.

Metabolites eluting from GC are ionized by Electron-impact (EI) or Chemical Ionization (CI;Watson, 1997; de Hoffmann & Stroobant, 2002). For metabolic analysis EI (Fiehn *et al.*, 2000b) is the most commonly used technique. In EI vaporized metabolites are impacted by a beam of electrons with sufficient energy to

fragment and ionize the molecule. The number of fragment ions that is produced of each metabolite is a function of the electron impact energy. The source is designed so that when the ions are formed they are pushed out from the source and into the mass analyzer. EI results in molecular ion fragmentation, which is of great importance for structural interpretation of the metabolites. In comparison with EI, CI is a much softer ionization technique, in which the ions are allowed to collide with reagent gas (often proton-rich) to form abundant adduct ions that contain the intact molecular species (Watson, 1997). This is advantageous for determining the molecular weight of metabolites. To identify compounds, commercially available databases of molecular ion fragmentation patterns of molecules, such as NIST (http://www.nist.gov/srd/nist1.htm; 19-August-2005), can be used. Unfortunately, the number of derivatized plant metabolites is limited and additional retention index information is incomplete. In addition to commercial libraries, in-house standard libraries (Bino *et al.*, 2004) have also been compiled containing spectra and corresponding retention indices. In the metabolomics science community databases have been compiled of collections of mass spectra and retention indices of frequently observed metabolites in plants (Wagner, Sefkow & Kopka, 2003; Schauer *et al.*, 2005).

Liquid Chromatography-Mass Spectrometry (LC/MS)

Instead of GC/MS, metabolites can be separated and detected by Liquid Chromatography (LC) coupled to atmospheric pressure ion sources (Herbert & Johnstone, 2002). ElectroSpray Interfacing/Ionization (ESI; Tolstikov & Fiehn, 2002) is the most widely used, more so than Atmospheric Pressure Chemical Ionization (APCI; Garratt *et al.*, 2005), for metabolic profiling. Unlike GC/MS, in LC/MS analyses the metabolites do not have to be volatile or possible to derivatize. After extraction and, if necessary, sample purification, the metabolites are separated by LC according to the differences in chemical properties of the metabolites present. Reverse phase chromatography is widely used in the field of metabolomics (Broeckling *et al.*, 2005). The separation of the extract depends on how the metabolites interact with the alkyl bonded spherical silica stationary phases. Many biologically important compounds do not separate easily on reversed-phase packing material, C_{18}, due to their high polarity. Hydrophilic interaction liquid chromatography (HILIC) methodology has also been used (Schlichtherle-Cerny, Affolter & Cerny, 2003; Tolstikov *et al.*, 2003). Recently, ultra high-pressure chromatography systems, such as UPLC™, have been developed to improve the separation efficiency (e.g. the number of components that can be separated/isolated from a mixture) of metabolites (Shen *et al.*, 2005; Wilson *et al.*, 2005a; Wilson *et al.*, 2005b). C_{18} monolithic silica capillary columns have been used in plant metabolomics to improve chromographic resolution (Tolstikov *et al.*, 2003). Both UPLC™ and monolithic columns improve peak separation, resulting in the ability to detect more peaks. In LC-ESI-MS (Fenn *et al.*, 1989) the LC effluent is transported through a capillary with a high voltage (2-5 kV). This leads to an electric field gradient forming on the water surface. The polarity of the voltage (positive or negative) chosen depends on the analyte. From the capillary tip a "Taylor cone" is generated, and at a certain point (when the Columbic repulsion of the surface charge is equal to the surface tension of the droplet) the droplet bursts and small charged droplets are formed that separate from it (Kebarle & Peschke, 2000). The droplets fly in atmosphere pressure towards the entrance of

the mass analyzer. In positive atmospheric electrospray ionisation an oxidation takes place at the spray tip and reduction on the counter metal plate. The ionization can be limited due to suppression effects by factors such as the presence of salts in the matrix of non-volatiles, e.g. various inorganic buffers (King *et al.*, 2000; Cech & Enke, 2001). Another limiting problem is that many analytes cannot be ionized or give a low ionization efficiency in ESI. As a rule of thumb, the analyte has to have a functional group that can be ionized, e.g. carboxyl or amine groups, and the ESI response increases with hydrophobicity (Cech & Enke, 2001). ESI responses can be improved by derivatization of the analytes (Okamoto, Takahashi & Doi, 1995; Leavens *et al.*, 2002; Nordström *et al.*, 2004). In comparison with electron impact in GC/MS both ESI and APCI fragmentations of molecular ions during ionization are much softer, and thus yield less information for mass interpretation. However, by running tandem MS (MS/MS) to provide fragmentation information (Huhman & Sumner, 2002; Tolstikov & Fiehn, 2002) and/or measuring accurate masses (Wilson *et al.*, 2005a) structural information can be generated for identification. Databases, such as those for GC/MS fragmentation patterns, are available, but only a few for LC/MS/MS. A problem with tandem mass spectra is that different instruments generate different spectra.

Mass Spectrometry

A mass spectrometer (MS) can be seen as an advanced balance that measures the weight of eluting metabolite fragments from the ion source (Watson, 1997; de Hoffmann & Stroobant, 2002). MS is a widely used approach for the identification and quantification of metabolites. In simplified terms, an MS consists of several parts: an ion source (in which ions are formed), a mass analyzer (which separates the formed ions according to their mass to charge ratio; m/z), a detector (which detects the separated ions) and a computer system (which controls the mass spectrometer and records the mass spectra). The different types of ion sources have been described above. The choice of mass analyzer in metabolomics depends on requirements with respect to scan speed, sensitivity, and selectivity for specific metabolites (reviewed by Dunn & Ellis, 2005). MS systems are mostly coupled after separation systems such as GC, LC or CE, but compounds can also be introduced using mass direct injection (Tohge *et al.*, 2005), Matrix Assisted Laser Desorption Ionization Time-of-flight Mass Spectrometry (MALDI; Edwards & Kennedy, 2005) and Desorption/Ionization on Silicon (DIOS). The most widely used mass analyzers in plant metabolomics are time-of-flight, quadrupole mass analyzers, and quadrupole ion trap mass systems. After the charged molecules have been produced they enter the mass analyzer. This is a low pressure region, evacuated by rotary pumps and turbo pumps. The quadrupole mass analyzer (de Hoffmann & Stroobant, 2002) is used for both GC/MS (Fiehn *et al.*, 2000b) and LC/MS (Idborg-Björkman *et al.*, 2003), and consists of two pairs of rods arranged orthogonally to each other. Each pair of rods employs a combination of direct-current (DC) and radiofrequency (RF) fields as a mass "filter". The "filter" is scanned by ramping the magnitude of the RF amplitude and DC voltages at a fixed ratio. For LC/MS quadrupole mass analyzers the quadrupoles are often aligned in series of three in order to run tandem MS (MS/MS) and thus provide fragmentation information. In a similar way, tandem MS can be used for metabolite identification with quadrupole ion trap mass spectrometers. TOF

(Mamyrin, 2001) is a simple method in comparison to other mass analyzers for mass measurements. In recent years TOF systems e.g. GC/TOFMS (Cook *et al.*, 2004) and MALDI/TOF (reviewed by Newton *et al.*, 2004) have been essential instruments for plant biological analyses. Ions from the ion sources are pushed into a low pressure flight tube where they drift and are separated by their masses; lighter ions travelling faster than heavier ions. The ions will have a kinetic energy distribution when they are pushed out in the tube, and to compensate for this a reflectron is placed at the end of the drift zone. This means that all ions will be focused and reflected back along the flight tube. Each ion will penetrate the field differently, depending on its kinetic energy. The time it takes for one ion to be detected from being pushed out corresponds to the mass of the ion. In contrast to quadrupole systems, which are scanning instruments, all ions will be detected in TOF.

Generating Analytical Protocols

To be able to compare the metabolome between plants a number of criteria have to be fulfilled: the biological variation should be kept low, the metabolites have to be chemically intact during analysis (i.e. the extraction should be non-destructive), the analysis should be global, quantitative and accurate, and the information should be interpretable. Bearing in mind that plants contain a wide spectrum of both low and high abundance metabolites a number of analytical aspects must be considered to fulfil the above criteria. This raises problems when choosing analytical protocols, as the optimum extraction conditions differ widely for different types of compounds. Ordinary analytical analysis is often focused on one analyte or a group of similar analytes. Unique target protocols are often used together with labelled internal standards for each analyte. In comparison with traditional quantification methods, a number of compromises have to be considered in metabolomics, where the aim is to analyze and identify as many metabolites as possible in as short a time as possible. Studies have been presented in which the analytical precision and biological variation in metabolomic investigations have been examined (Shurubor *et al.*, 2005).

Extraction

To stop metabolic processes and freeze the metabolome the plant samples can be freeze-dried (Le Gall *et al.*, 2003b) or rapidly frozen in liquid nitrogen (Cook *et al.*, 2004). After sampling, the samples can be stored for a couple of days or weeks at -80°C to maintain stability until extraction. However, there are limitations in the time that samples can be stored in a freezer. To extract plant metabolites efficiently from plant tissues, the tissue has to be homogenized properly first. The efficiency with which the solvent can penetrate the tissue influences the length of time required for solvent extraction and the degree of homogenization. Various techniques have been applied to accelerate this process, e.g. grinding with a mortar and pestle together with liquid nitrogen, milling in vibration mills with chilled holders, homogenization with a metal pestle connected to an electric drill (Edlund *et al.*, 1995) and ultra turrax devices (Orth, Rentel & Schmidt, 1999). A common way to extract metabolites is to shake the homogenized plant tissue at low or high temperatures in organic solvents, or mixtures of solvents (Johansen, Glitso & Knudsen, 1996; Streeter & Strimbu, 1998; Cook *et al.*, 2004; Broeckling *et al.*, 2005). For polar metabolites, methanol, ethanol

and water are often used, while for more lipophilic compounds chloroform is the most commonly applied solvent. Other extraction techniques include supercritical fluid extraction (SFE; Huie, 2002) microwave-assisted extraction (MAE; Barclay, Bonner & Hamilton, 1997; Kaufmann & Christen, 2002), subcritical water extraction (SWE; Gamiz-Gracia & de Castro, 2000; Ozel *et al.*, 2005) and pressurized liquid extraction (PLE; Benthin, Danz & Hamburger, 1999; Ong, 2002), but these methods have not been widely adopted, as yet, to extract metabolites from plant tissues. In metabolomic analyses, the goal is to analyze as many metabolites as possible in a single analysis (e.g. single GC/MS run), so the extract is not usually purified, in contrast to routine procedures for analyzing specific metabolites in complex matrices where, for instance solid phase extraction (SPE) is commonly used. However, it is common to divide metabolomic extracts into polar and lipophilic fractions by solvent partitioning (Broeckling *et al.*, 2005; Desbrosses, Kopka & Udvardi, 2005).

Derivatization for GC/MS

The most common way to derivatize polar compounds containing functional groups, such as –OH, –SH or –NH groups, is to add a trimethylsilyl (TMS) group, and form TMS-ethers, TMS-sulfides or TMS-amines, respectively (Pierce, 1968; Blau & Halket, 1993). TMS-ethers of mono- and di-saccharides are easily prepared and separated chromatographically, but TMS-derivatization of monosaccharides often results in the formation of multiple peaks since reducing sugars occur in solution as mixtures of different anomers. TMS-derivatization usually results in five tautomeric forms of the reducing sugars (Curtius, Muller & Völlmin, 1968; Asres & Perreault, 1997). Due to the equilibrium shifts and separation problems entailed, this causes major problems for the identification and quantification of complex mixtures of carbohydrates in plant tissues. However, by converting the aldehyde and keto-groups into oximes using hydroxyamines or alkoxyamines before forming TMS-ethers, the number of tautomeric forms can be reduced, due to the limited rotation along the C=N bond, resulting only in the formation of *syn* and *anti* forms (Fiehn *et al.*, 2000b; Figure 1). Several different reagents can be used both for oximation and silylation. The choice of reagents depends on the reaction efficiency, to ensure that as few metabolites as possible remain underivatized. One drawback associated with using silylation derivates is the formation of unexpected side products, artefacts (Little, 1999). MSTFA (N-methyl-N-trimethylsilyltrifluoroacetamide) derivatization in combination with oximation is one of the most widely used procedures for sugars and plant metabolites (Fiehn *et al.*, 2000a; Duran *et al.*, 2003). Other alternatives are derivatization by BSTFA (N,O-bistrimethylsilyltrifluoroacetamide) and BSA (N,O-bistrimethylsilylacetamide), but MSTFA is more volatile and thus more suitable for direct GC analysis (Walhout & Pierce, 1968). It has also been shown that different catalyzing compounds and reagents, such as potassium acetate, pyridine, TMCS and TMBS, can enhance the silylation power of the silylation reagents (Evershed, 1993). Silylation efficiency can also affect the choice of solvent (Walhout & Pierce, 1968; Evershed, 1993; Adams *et al.*, 1999). To improve the identification and structural information obtained, N-methyl-N-tert-butyldimethylsilyltrifluoro acetamide (MTBSTFA) has been used for fragment assignment (Fiehn *et al.*, 2000b).

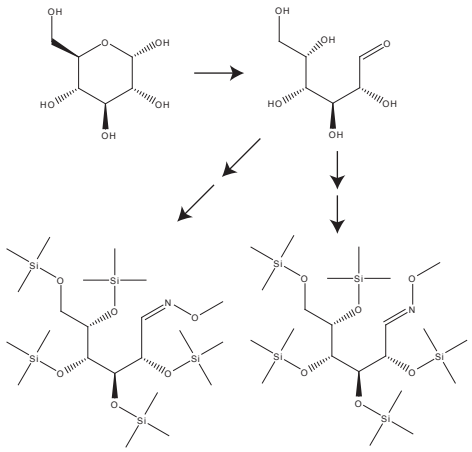


Figure 1. Principle of derivatization for metabolomic analysis by GC/MS. Aldehyde and keto-groups are converted into oximes using methoxyamine followed by conversion of hydroxyl groups into trimethylsilyl (TMS) groups. The example shows derivatization of glucose. Two tautomeric forms, *syn* and *anti*, are formed due to rotation along the C=N bond.

Design of Experiment

In biology and analytical chemistry there is a need to perform experiments in a systematic way. To generate a protocol for global analysis of metabolites or to design a plant experiment for metabolic profiling, a number of factors (e.g. biological treatments, genotypes, temperature, amount of reagents and instrumental setups) can be identified which can affect the responses (e.g. increased growth, changes in internode length, hormone status, metabolite response, reproducibility or yield). The goal for the analytical systems is often to find the settings that maximize the response and reproducibility. The traditional way to investigate and find optima in an experimental domain (the experimental 'area' that is defined by the variation of the experimental variables) is to Change One Separate factor at a Time, i.e. the COST approach. Finding true optima is not straightforward with this approach, as it is inefficient (requiring unnecessarily large numbers of runs), it ignores interactions, it generates knowledge relatively slowly and does not map the experimental space. Design of experiment (DOE, Lundstedt *et al.*, 1998; Carlson & Carlson, 2005; Riter *et al.*, 2005) is a procedure where variation is introduced systematically to the experimental domain and the effects of these factors are analysed using regression models. In contrast to the COST approach DOE allows the causal effect of each factor in the experimental domain to be elucidated in relatively few experiments in a systematic way. A factor is here defined as an experimental variable that can be changed independently of the others. By measuring the effect of one experimental variable at different levels of another variable it is also possible to estimate their interactive effects. If the combinations of k factors are investigated at two levels, a factorial design will consist of 2^k experiments (Lundstedt *et al.*, 1998). In a factorial design the influence of all experimental variables, factors, and interaction effects on the response or responses are investigated. For example if the effects of two factors, time and temperature, on derivatization are to be investigated at two levels (e.g. 24h /48h and 20/60°C) four experiments (2^2 design = 4 experiments) must be performed to fulfil the criteria of independent experiments and to estimate interactions between the factors The experiments are: A (24h/20°C), B (48h/20°C), C (24h/60°C) and D (48h/60°C). To estimate the experimental error a number of experiments are

repeated, preferably in the centre of the experimental domain.

The DOE strategy can be divided into the following main parts: problem formulation, planning of experiments and measurements of the responses according to the design, evaluation and interpretation of the model (Figure 2). In problem formulation the questions "What are the purposes?" and "What are the objectives?" are addressed, and the answers decide the design setup and number of experiments. The outcome from the experiments are responses that are commonly modelled using Ordinary Least Squares (OLS; Martens & Naes, 1992) or Partial Least Squares Regression (PLS; Höskuldsson, 1995; Wold, Sjöström & Eriksson, 2001) OLS can be used when the number of experimental factors is equal to or fewer than the number of experiments and the design factors (columns) are uncorrelated. If the different design factors are highly correlated with each other, for example in mixture designs, PLS can be used as one of many alternative regression methods (Antti *et al.*, 2004). The model setup is $\mathbf{y} = b_0+b_1x_1+b_2x_2……..+b_{12}x_1x_2+\mathbf{f}$, where b represents the unknown estimate.

Depending on the design setup, different equations can be applied to give an approximation of the response surface. These equations may contain linear, interaction and square terms in attempts to fit the true surface of the domain. From the model important factors that influence the results can be used for interpretation. The model can also be used to make predictions to validate the model. The statistical validity of the model is evaluated by using cross-validation (Wold, 1978) and to compare variation in the model and the experimental error (Lundstedt *et al.*, 1998). After the experiments and interpretation a new round of DOE can be performed, in which either the complexity of the system is reduced or the values for specific variables are selected. The settings and the number of levels for the different factors are selected according to the question addressed, the complexity of the system and the number of suitable experiments. Designs can also be divided into screening and response surface modelling (RSM) types. For screening investigations only the effects of the experimental variables and interactions are estimated. After screening, the goal of an investigation is usually to approximate the response by a quadratic polynomial. Alternative designs, such as determinant-optimal designs (D-optimal designs) can be used for unsymmetrical designs (de Aguiar *et al.*, 1995) and for mixture designs (Lundstedt *et al.*, 1998) where the design variables cannot be varied independently of each other. D-optimal designs maximize the experimental space for a selected number of experiments for defined model equations.

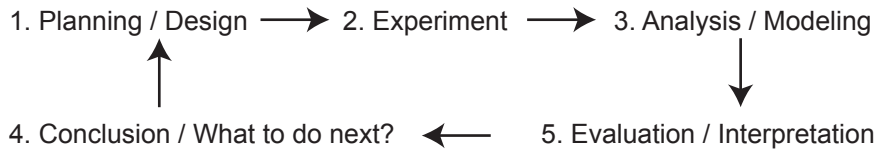


Figure 2. An overview of Design of Experiment steps.

Multivariate Analysis

Experiments are performed according to the biological or analytical question to be addressed, and the DOE strategy can be used to introduce variation between samples systematically, Figure 2. For biological samples analyzed by modern analytical

techniques, such as LC/MS, GC/MS and NMR, the number of responses is often large (Figure 3). Due to the complexity of the samples the different variables are often correlated. For example, environmental changes can affect many metabolites belonging to the same metabolic pathway in a plant. Often, the goal of data mining is to find structures in experimental information and to describe it in an interpretable and simple way. The traditional approach is to consider one variable at a time, but chance correlations may give false outcomes. To help understand the measurement information it can be divided into an interpretable model and noise. This can be done in numerous ways, one of which is to explore the underlying data structure for a set of samples by reducing the number of variables to a few independent principal components (PCs) or latent variables (Kvalheim, 1992). The different latent variables are built up by linear combination of different variables describing similar variation. According to the experimental setup the latent variables can be used to generate models to describe this variation. Depending on the type of information to be explored from the data table different latent variables can be calculated. To obtain an overview of a data table, to detect clusters, patterns and trends between the samples, and to identify abnormal samples Principal Component Analysis (PCA; Jackson, 1991; Höskuldsson, 1995) can be used. Partial Least Squares Regression (PLS) is the regression analogy of PCA, where the relationship between data tables is sought. The use of multivariate calibration in analytical chemistry is reviewed in more detail by Bro (2003). Other methods used to model variation in metabolomics include HCA (hierarchical cluster analysis; Sumner, Mendes & Dixon, 2003), discriminate analysis (DA; Allen *et al.*, 2003), correlative network analysis (Steuer *et al.*, 2003), neural networks (Taylor *et al.*, 2002) and Genetic algorithms (Goodacre, 2005).

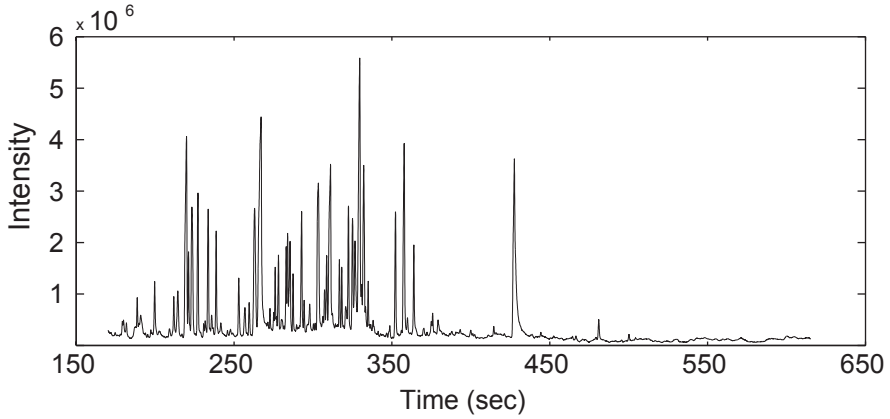


Figure 3. Total ion current chromatogram (TIC) from a typical analysis of a methoxymated and trimethylsilyl derivatized extract from *Arabidopsis*.

Annotation

PCA and PLS are best described using linear algebra and vector-matrix notation. Bold capital letters (**X**) are used for matrices. Small bold characters (**p**) are used for column vectors. Transposition is also used to make row vectors from column vectors and vice versa. ^{T} superscripts denote transposition, and $^{-1}$ the inverse of a matrix. Small italic letters (*k*) are used for scalars. ||**w**|| represents the length of vector **w**.

Principal Component Analysis (PCA)

The metabolic information can be represented as an **X** matrix where each row represents a sample (N) and each column (K) represents an instrumental or metabolite response. Instead of describing the information as a data table the information can be visualized in a multidimensional space where each sample is represented as a point in a K-dimensional space. By doing so a swarm of samples can be projected from the K-dimensional space down to fewer dimensional hyper-planes, which can be regarded as different two-dimensional windows (score-plots). The score vectors will give a good approximation of the location of the different samples in the K-dimensional space. Each hyper-plane's direction in the K-dimensional space corresponds to loading. For example, if three responses have been measured for six samples (Table 1), each sample can be represented as a point in a three-dimensional coordinate system. The samples are then mean centred (Table 2), and introduced into a three dimensional space (Figure 4). Centering is performed by subtracting the mean value for each variable from each corresponding variable response.

Table 1. Three metabolite responses for six plant samples.

Sample	Metabolite 1	Metabolite 2	Metabolite 3
Sample 1	14	15	17
Sample 2	10	17	16
Sample 3	13	13	15
Sample 4	13	4	6
Sample 5	12	3	4
Sample 6	10	2	5
Mean	12	9	10.5

Table 2. The metabolite information from Table 1 after subtraction of the mean value for each column from each row. Score, $\mathbf{t}_1$ and $\mathbf{t}_2$, and loading values, $\mathbf{p}_1$ and $\mathbf{p}_2$, are calculated for the matrix. The values are rounded.

Sample	Metabolite 1	Metabolite 2	Metabolite 3	$\mathbf{t}_1$	$\mathbf{t}_2$
Sample 1	2.0	6.0	6.5	**8.9**	**1.8**
Sample 2	-2.0	8.0	5.5	**9.5**	**-2.7**
Sample 3	1.0	4.0	4.5	**6.0**	**0.9**
Sample 4	1.0	-5.0	-4.5	**-6.7**	**1.2**
Sample 5	0.0	-6.0	-6.5	**-8.8**	**0.1**
Sample 6	-2.0	-7.0	-5.5	**-9.0**	**-1.4**
$\mathbf{p}_1$	**0.04**	**0.74**	**0.67**		
$\mathbf{p}_2$	**0.95**	**-0.24**	**0.21**		

PCA will reduce the dimensionality of the multidimensional space and the data matrix by introducing a number of new orthogonal linear independent vectors (latent variables). This is done by first introducing one PC (Principal Component) $\mathbf{t}_1$ describing the most variance in the K-dimensional space and projecting each sample down onto the new vector, Figure 4.

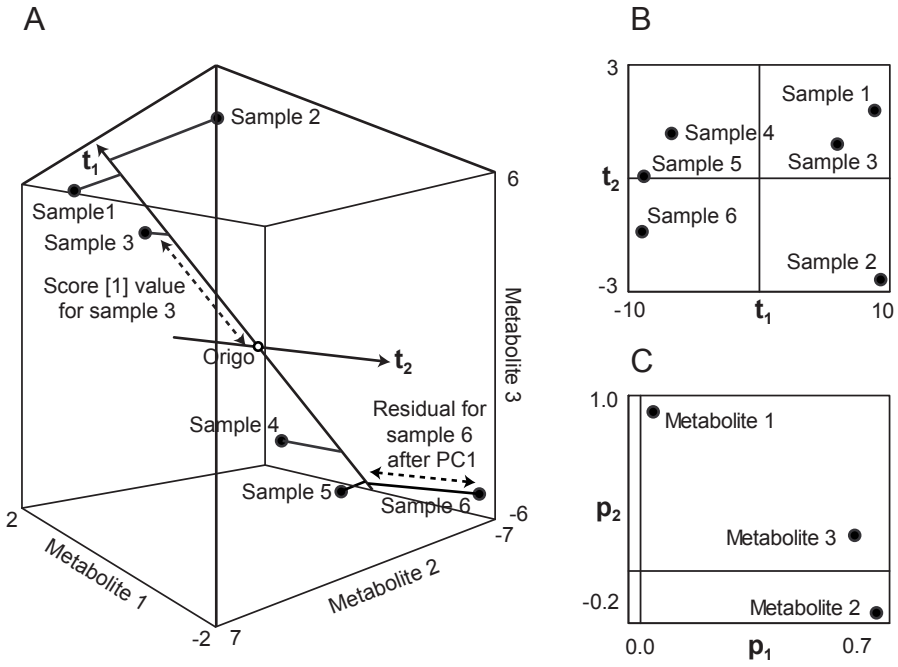


Figure 4. A: A projection of the samples in Table 2. Score vectors $\mathbf{t}_1$ and $\mathbf{t}_2$ describe the largest and second largest variation in the data. The score value for each sample is the distance between the projection and the mean centre of the data swarm. The score values describe the relation between samples. Loading **P** describes the importance of the different variables for describing the variation in the PCs. The cosine of the angle between the principal component directions and each of the original coordinate axes corresponds to loading **P**. B: A projection plot of the three-dimensional space down to a two-dimensional hyper plane ($\mathbf{t}_1/\mathbf{t}_2$) C: In the first latent variable metabolites 2 and 3 are the most important (describing the most variation), as can be seen in the first loading $\mathbf{p}_1$ (Table 2).

Decomposition of a mean centred **X** matrix to the scores, loadings and residuals can be written:

(1) $\mathbf{X} = \mathbf{TP}^T + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots \mathbf{t}_a\mathbf{p}_a^T + \mathbf{E}$

The data reduction is accomplished by neglecting unimportant directions where the sample variation is insignificant. This is repeated until no significant direction in the K-dimensional is left, i.e. the residual. The maximum number of components (a) is the same as the number of variables. The number of significant PCs can be estimated by a number of methods, such as calculating the size of eigenvalues (Jackson, 1991) or cross-validation. After all significant variation in **X** has been described by the PCA model the remaining variation, the residual, is non-systematic and represents the distance between each point in K-space and its point on the plane.

Partial Least Squares Regression (PLS)

PCA is an unsupervised method for which no additional information about the data is needed. The data describing the most variation in the X matrix are projected down to hyper planes. In contrast to PCA, PLS is a supervised method where additional

information is used to find information (**y**) in the **X**-matrix. PLS will find the linear relation between the **X** matrix and an external data vector (**y**). Like PCA, PLS is designed to find the latent structure in the **X-**matrix. PLS maximizes the covariance between **X** and **y**. The **X** matrix and **y** vector are decomposed in a similar way as for PCA:

(2) $\mathbf{X} = \mathbf{TP}^T + \mathbf{E} = \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1} + \mathbf{E}$
(3) $\mathbf{y} = \mathbf{Tc}^T + \mathbf{f} = \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{c}^T + \mathbf{f} = \mathbf{Xb} + \mathbf{f}$

where: **T** is the score matrix of **X**; **P** and **c** are the loading matrices of **X** and **y**, respectively; **W** is the weight matrix of **X**; **E** and **f** are the residual matrices for **X** and **y**, respectively. The number of latent variables used in the PLS model depends on the predictability and is estimated using, for example, cross-validation. The regression coefficient for the PLS model can be expressed as:

(4) $\mathbf{b} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{c}^T$

If the criteria for OLS are fulfilled for the **X** matrix, the outcome of the PLS model will be the same as for OLS. The main advantage of PLS is that the information from the data can be correlated with response data rather than simply describing the variation in the **X**-matrix, as PCA does. The score values and loading values can be explored in a similar way as in PCA to interpret the biological implications.

I have here only described PLS for the case of a single **y** vector, but the algorithm can also be used for regression between two or more **y** (**Y**) vectors and the **X**-matrix. The use of PLS for two blocks is discussed in further detail by Wold, Sjöström & Eriksson (2001) and Trygg & Wold (2003). A similar prediction method to PLS is Orthogonal projection on latent structure (O-PLS) developed by Trygg & Wold (2002, 2003). For O-PLS the variation in **X** is divided into three parts instead of two, as in PLS. For a single **y** the **X** matrix is decomposed according to:

(5) $\mathbf{X} = \mathbf{t}_p\mathbf{p}_p^T + \mathbf{T}_o\mathbf{P}_o^T + \mathbf{E}$
(6) $\mathbf{y} = \mathbf{t}_p\mathbf{c}_p^T + \mathbf{f}$

The first part of the variation in **X** is used to predict the variation in **y** ($\mathbf{t}_p\mathbf{p}_p^T$), the second part contains so-called structured noise, variation that is orthogonal to **y** ($\mathbf{T}_o\mathbf{P}_o^T$), and the third part is residual variance. For PLS the variation is only separated into two parts: the variation used to model the variation in **y** and residual variance. The fraction of the sum of squares, R^2X, of all the **X**'s explained by the current component can be divided for O-PLS into R^2X_{corr} (the variation used to predict **y**) and R^2X_{yo} (**y** orthogonal variation). O-PLS gives similar predictions of **y** to PLS, but the interpretation of the models is improved because the structured noise in the model is separated from the variation describing the variation in **y**.

Deconvolution

Hyphenated chromatography and mass spectrometry systems, including GC/MS and LC/MS, are commonly used to quantify, identify and screen for new metabolites in modern biology. To fulfil the requirements for high throughput, analyses are often short and the chromatograms often complex, containing hundreds of completely or partly overlapping peaks. To obtain the chromatographic and spectral profiles, especially to distinguish between overlapping peaks, curve resolution (commonly

named deconvolution) methods have been developed. Peak alignment and other pre-treatments of other types of data, for instance GC/MS data, are not discussed further in this thesis. For coverage of these topics, see reviews by, for example, Fraga, Prazen & Synovec (2001) and Duran *et al.* (2003).

Curve resolution

Samples from hyphenated techniques generate a matrix, **X**, where each row, N, represents spectra measured at one time point and each column, K, represents chromatographic profiles for a signal (e.g. a mass channel or wavelength). Deconvolution methods decompose the two-way signals into a number of unique peaks and spectra. According to the Lambert-Beer Law the matrix can be decomposed into, and expressed as:

(7) $\mathbf{X}=\mathbf{C}\mathbf{S}^{T}+\mathbf{E}=\mathbf{c}_1\mathbf{s}_1^{T}+\mathbf{c}_2\mathbf{s}_2^{T}+\ldots\ \mathbf{c}_a\mathbf{s}_a^{T}+\mathbf{E}$

This is done for the number of profiles (a) in the system. **C** represents the unique chromatographic profiles, **S** the corresponding spectral profiles and **E** the residual. This decomposition can be performed using, for example, PCA. Consider a sample, **X,** containing three components (GC/MS analysis of TMS-derivatized [$^2H_4$]-succinic acid, succinic acid and norleucine) with unique chromatographic (Figure 5) and mass spectral profiles (Figure 6). All masses that are measured by hyphenated techniques can be used to explore the three eluting peaks.

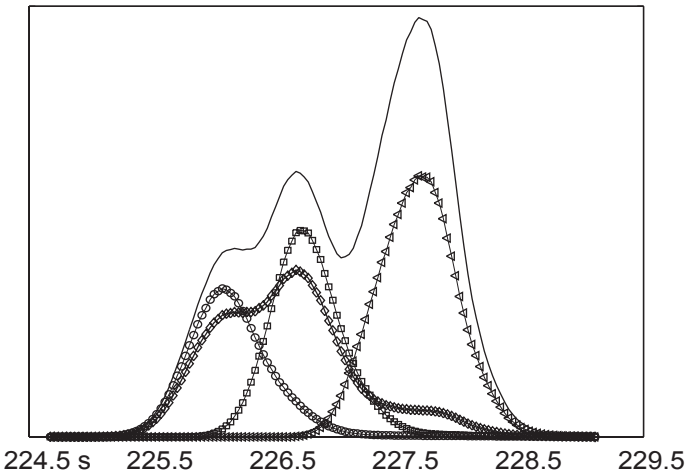


Figure 5. Three eluting peaks, where the black lines correspond to TIC, the circles to m/z 251 ([$^2H_4$]-succinic acid-TMS), the squares to m/z 247 (succinic acid-TMS), the triangles to m/z 158 (norleucine-TMS) and the diamonds to m/z 147. Both TIC and m/z 147 are scaled down to fit the plot.

By using PCA and projecting down the swarm of samples, in this case retention times, an overview of the analytical information can be obtained by exploring the score values (Figure 7A). The most important variables, m/z channels, describing the variation in the different latent variables can be seen in the corresponding loading plot (Figure 7B). By looking in the direction of the time of the peak maximum

of each eluting peak a good approximation of the component mass spectra can be estimated. The goal in deconvolution is to find directions, e.g. using PCA, in this multidimensional space that give a solution to equation 7.

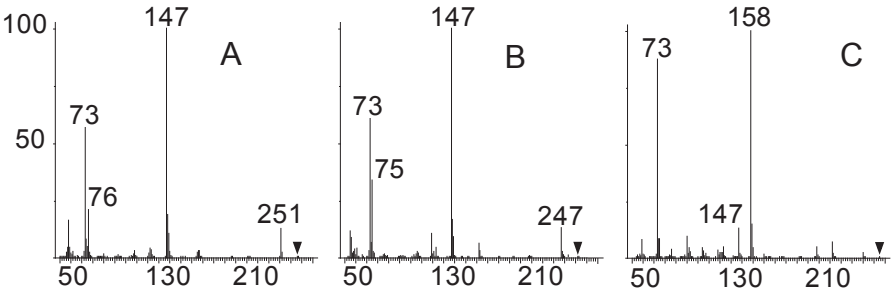


Figure 6. Pure mass spectra of TMS-derivatized [$^2H_4$]-succinic acid (A), succinic acid (B) and norleucine (C). The triangles represent the mass of the molecule ion.

No *a priori* knowledge is needed about the bilinear decomposition of data except that the pure chromatographic and spectral profiles should only contain positive values and the chromatographic profiles should only contain one peak (unimodality). The constraint for PCA is that each PC has to be orthogonal to each other, so each PC needs to be rotated to give a good solution for **C** and **S**.

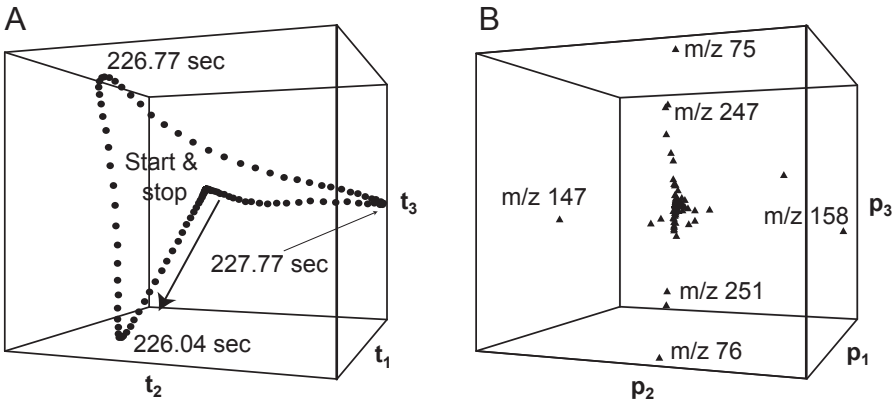


Figure 7. A: Score plot for $\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{t}_3$ for the three eluting peaks in Figure 5. Projection of retention times (samples) down from a 750-dimensional m/z space to three principal component dimensions. B: Loading plot for $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{p}_3$. The plot describes the most important m/z describing the variation in A.

Methods for the solution of the composition can be divided into unique and rational resolution methods (Jiang, Liang & Ozaki, 2004). The unique methods, such as heuristic evolving latent projections (HELP; Kvalheim & Liang, 1992) and orthogonal projections (OP; Liang & Kvalheim, 1994) try to pick regions in the chromatogram that are unique for some single compound and use them to decompose the **X** matrix. The rational resolution methods, such as alternating regression (AR; Karjalainen, 1989), iterative target transformation factor analysis (ITTFA; Gemperline, 1984) and Simple to use interactive self-modelling mixture analysis (SIMPLISMA; Windig & Guilment, 1991), may produce sets of possible solutions and depend on the

similarity between spectral and chromatographic profiles. Provided the correlation and collinearity between the profiles are not too strong, the solution will be a good approximation of the true profiles. The curve resolution methods can also be divided into non-iterative, iterative, and hybrid approaches (Liang & Kvalheim, 2001). The non-iterative methods, such as OP and HELP, involve rank analysis of evolving matrices. The main disadvantage with these methods is that they are very difficult to automate, due to the need to define analyte elution windows by local rank analysis. The iterative methods all define start profiles, but the procedures for selecting initial estimates and resolution differ amongst them. Examples of iterative methods include ITTFA and AR. For the iterative methods, it is essential to estimate the chemical ranks of profiles correctly to generate the right solution. These approaches are described in detail in a number of review papers (Toft, 1995; Sanchez *et al.*, 1996; Grande & Manne, 2000). Originally the AR algorithm used random numbers as a starting estimate for **S**. **C** is calculated using least squares:

(8) $\mathbf{C}=\mathbf{XS}(\mathbf{S}^T\mathbf{S})^{-1}$

C is corrected according to constraints (unimodality and non-negativity):

(9) $\mathbf{S}=\mathbf{X}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}$

S is corrected to non-negativity constraints. The new spectral estimates are corrected according to constraints. New chromatographic and spectral profiles are calculated until convergence by iteration between equations 8 and 9. The disadvantage is, generally, that high-quality results require a good choice of starting vectors. Hybrid methods, such as automatic window factor analysis (AWFA; Malinowski, 1996) and Gentle (Manne & Grande, 2000), start from a set of key spectra from which concentration and spectral profiles are estimated.

Generally in deconvolution, samples must be resolved separately and the estimated spectral profiles of all samples must be carefully checked in order to obtain reliable mass spectra and peak areas. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS, Tauler, 1995) is a method where the **X** matrix is instead built up by all samples to be compared. The data can be regarded as a data cube of size N*K*L, where N = the number of samples, K = the number of time points (scans) and L = the number of m/z channels. The cubes are then unfolded to form a data matrix **X** of size (N*K) * L. The matrix is decomposed in a similar way (Tauler, 1995) as for AR and the outcome will be **C** of size (N*K) * the number of resolved components (R; i.e. the number of components with a common mass spectrum) and **S** of size L*R. Each sample will yield a chromatographic profile for each resolved component. This procedure has been used, for example, to deconvolute data obtained in liquid chromatography with diode array detection (LC-DAD; Tauler, Lacorte & Barcelo, 1996; Pere-Trepat *et al.*, 2004) and LC-NMR analyses (Bezemer & Rutan, 2001).

Applications and Software for Deconvolution

In the field of metabolomics automatic peak detection and mass spectrum deconvolution for GC/MS can be performed using instrument-specific software such as ChromaTOF™ software (Leco Corp., St Joseph, MI, USA). Automatic curve resolution of chromatograms can also be applied with a high degree of success,

e.g. by using the freely available AMDIS software (http://chemdata.nist.gov/mass-spc/amdis/, 19-August-2005; Halket *et al.*, 1999). Gentle has been successfully used for automatic data processing of GC/MS (Eide *et al.*, 2001) and LC/MS data (Idborg-Björkman *et al.*, 2003). In addition, Micromass MarkerLynx applications manager (Lenz *et al.*, 2004; Wilson *et al.*, 2005a) and metabAlign software (Tolstikov *et al.*, 2003; Vorst *et al.*, 2005) have been used to filter out differences from raw data files between metabolic samples. Chromatographic alignment of raw GC/MS data can be carried out before deconvolution as metabolomic data, spectral formatting, alignment and conversion tools (MSFACTs; Duran *et al.*, 2003).

Gibberellin Interactions with Auxin

Gibberellins (GAs) are endogenous plant growth regulators that control numerous aspects of plant growth and development, such as stem elongation, leaf shape, root and fruit growth, flowering and flower development (Davies, 2004; Sun & Gubler, 2004; Tyler *et al.*, 2004). The importance of GAs has been shown by studying mutants with reduced biosynthesis of bioactive GAs. Such mutants are generally dwarfed, with short internodes, small leaves, delayed flowering and varying degrees of male sterility. The phenotype of these mutants, e.g. *ga1-3* (Silverstone *et al.*, 1997), can be restored to wild-type-like by treatment with active GAs (King, Moritz & Harberd, 2001)

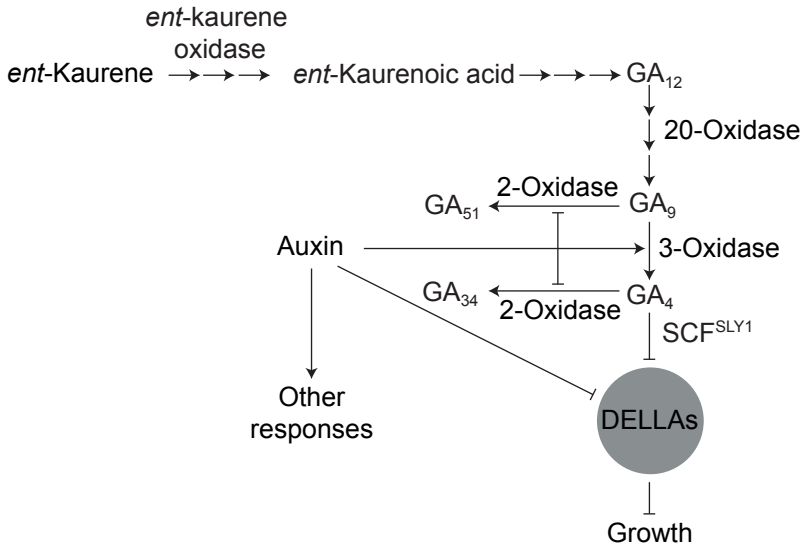


Figure 8. GA biosynthesis, auxin crosstalk and DELLA proteins. The active GA_4 causes degradation of the DELLA protein via the SCF^{SLY1} E3 ubiquitin ligase complex. Auxin works as a putative regulator of GA biosynthesis. Adapted from (Swain & Singh, 2005)

In recent years our understanding of GA biosynthesis and signalling has increased substantially through the identification of many genes involved in these processes. The biosynthesis of GAs can be divided into three stages, (i) the formation of *ent*-kaurene, (ii) the conversion of *ent*-kaurene to GA_{12} and (iii) the formation and deactivation of the bioactive GAs, GA_1 and GA_4 (Hedden & Phillips, 2000; Alvey & Harberd,

2005; Swain & Singh, 2005; Figure 8). Studies in *Arabidopsis*, rice and barley have also identified several positive and negative regulators of GA signalling pathways, all involved in regulating GA responsiveness during development (reviewed in Olszewski, Sun & Gubler, 2002 and Gomi & Matsuoka, 2003). The most intensively studied GA signalling components are the DELLA proteins (Sun & Gubler, 2004; Alvey & Harberd, 2005; Fleet & Sun, 2005). DELLA proteins are highly conserved among different plant species and belong to the GRAS protein family (Bolle, 2004). Five DELLA proteins have been identified in *Arabidopsis*; GA-Insensitive (GAI), Repressor of *ga1-3* (RGA), RGA-like1 (RGL1), RGL-2 and RGL-3. The different proteins are putative transcription factors, and are all thought to encode negative regulators of GA responses, with different roles during the life cycle of the plant; for instance RGA and GAI are the major repressors during vegetative growth (Dill & Sun, 2001; King, Moritz & Harberd, 2001; Bolle, 2004; Tyler *et al.*, 2004) and RGL2 has a role during seed germination (Lee *et al.*, 2002). DELLA proteins are rapidly degraded by the ubiquitin-proteasome pathway in response to active GA (Sun & Gubler, 2004). The SCF protein that is involved in the degradation is suggested to have a subunit called SLY1 (McGinnis *et al.*, 2003) and has been indicated to have a key role in GA responses.

In recent years several investigations have found evidence of cross-talk between GA biosynthesis and auxin levels. For example, reductions in IAA levels, caused by removing the apical bud in pea resulted in the down-regulation of GA3ox expression and lower levels of the bioactive GA_4 in studies by Ross *et al.* (2000). However the GA and IAA levels were restored by IAA applications. Similar results have also been found in tobacco, although the main step in GA biosynthesis affected after decapitation was the GA_{19} to GA_{20} transformation (Wolbang & Ross, 2001). Interestingly, recent results of investigations of GA signalling mutants have shown further interactions between GA and auxin-stimulated growth. There is now evidence that auxin stimulated growth can act via DELLA protein degradation (Fu & Harberd, 2003). These results suggest that DELLA proteins are general inhibitors of growth, and that GAs act entirely through the DELLA proteins, whereas auxin can also regulate growth and development independently of the DELLA proteins.

Objectives

The overall goal of the work described in this thesis was to develop high through-put MS and data comparison techniques for plant metabolomics. This methodology was then applied in a study of *Arabidopsis* mutants in order to extend our understanding of growth regulation by the plant hormone gibberellin (GA) and crosstalk between GAs and auxin.

The main objectives were to:

-Develop a fast and robust method for routine analysis of plant metabolite patterns using GC/MS.

-Use multivariate techniques to improve and accelerate the comparison of samples from high through-put MS metabolomic analyses.

-Generate methods to improve the identification and quantification of plant metabolomic data to help generate biologically interpretable results in a high through-put manner.

-Investigate the crosstalk between GAs and IAA by studying auxin levels and metabolite patterns in *Arabidopsis* mutants lacking GAs and/or parts of the GA-signalling pathways after GA application.

Experimental

Derivatization Design

To identify factors affecting the derivatization of plant metabolites twelve standard compounds were selected (D-ribose, alpha-ketoglutaric acid, glucosamine, D-fructose 6-phosphate, sucrose, N-acetyl-D-(+)-glucosamine, oxalic acid, L-proline, thymine, stearic acid, cholesterol and glycerol monostearate) that commonly occur as endogenous compounds in plants, and methyl octadecanoate as an internal standard (Paper I). The different metabolites and the internal standard, methyl octadecanoate, were dissolved and three µg of each standard compound were added to GC-vials. The solvents were evaporated in a Speed-Vac concentrator centrifuge (Savant Instruments, Farmingdale, NY, USA). Factors investigated were different temperatures, times and concentrations of different solvents during methyloximation and trimethylsilylation. The derivatization procedure used is shown in Figure 9. A D-optimal design was generated (Johnson & Nachtsheim, 1983; Dumouchel & Jones, 1994) to maximize the experimental space for 34 experiments for a model with linear terms for all of the factors, and interaction terms for all factors except the mixture factors. To each vial 20 µL of methoxyamine hydrochloride (20 mg/ml) was added, dissolved in different mixtures of pyridine/chloroform (Table 3) and vortex-mixed for 5 min. The vials were heated at 20, 40, 60 °C for 1, 9 or 17 h then 10 µL of MSTFA plus an additional 30 µL of MSTFA, heptane and acetonitrile mixture was added (Table 3). After vortex-mixing the mixtures were heated at 20, 40 or 60 °C for 30 min. All samples were analyzed in randomized order and the first sample

was analyzed by GC/MS 30 minutes after the silylation regents were added. To avoid errors due to chromatographic variations when different solvents were used, or variations in injection precision, all of the integrated peak areas were normalized by dividing them by the (underivatized) peak area for methyl octadecanoate. To correct for changes in retention time caused by the different solvents, the number of peaks estimated by the ChromaTOF software was restricted to peaks appearing after the first eluted metabolite, oxalic acid.

Table 3. Factor settings for a D-optimal design generated for 34 experiments to investigate derivatization using 12 standard compounds. The values are the minimum and maximum values of the design factors defining the experimental domain. *Marks dependent factors, the sum was always 40 µL.

Nr	Factor	Settings
1	Oximation, solvent	50/50% or 0/100% (chloroform/pyridine)
2	Oximation, temp	20 or 60°C
3	Oximation, duration	1 or 17 hour
4	Silylation, temp	20 or 60°C
5	Silylation, TMCS	1 or 10%
6	Silylation, solvent 1*	0 to 30 µL Heptane
7	Silylation, solvent 2*	0 to 30 µL ACN
8	Silylation, reagent *	10 to 40 µL MSFTA
	Replicate points	25% chloroform, 75% pyridine, 40°C, 9 hour, 40°C, 5.5% TMSC and a mixture of 10µL heptane, 10µL ACN and 20 µL MSFTA.

Extraction Design

In the study described in Paper I rosette leaves from ten three-week-old *Arabidopsis thaliana* (Co) plants grown under short day conditions in soil, were harvested and pooled, immediately thereafter frozen in liquid nitrogen and homogenized using a mortar and pestle. Approximately 22 mg of each sample was transferred to a 1.5 mL Eppendorf tube and frozen at -80°C overnight. Using a MM 301 Vibration Mill (Retsch GmbH & Co. KG, Haan, Germany) the samples were extracted at a frequency of 30 Hz, with 3 mm tungsten carbide beads. Seven factors affecting the extraction and derivatization (Figure 9) of metabolites from plant tissue were investigated according to a 2^{7-1} fractional factorial design (Carlson & Carlson, 2005). The factors were varied according to Table 4. In order to extract as many different metabolites as possible, a soluble mixture of chloroform, water and methanol was chosen. According to the design, 200 µL of chloroform, 200 µL of water and 600 µL of methanol were added either together or separately in the given order. Before extraction, isotope-labelled reference compounds, [$^{2}H_{7}$]-cholesterol, [$^{13}C_{3}$]-myristic acid and [$^{13}C_{4}$]-hexadecanoic acid, together with methyl octadecanoate, were added to the chloroform buffer. The rest of the labelled reference compounds ([$^{2}H_{4}$]-succinic acid, [$^{13}C_{5}$, ^{15}N]-glutamic acid, [$^{13}C_{5}$]-proline, [$^{13}C_{4}$]-disodium alfa-ketoglutarate, [$^{13}C_{12}$]-sucrose, [$^{2}H_{4}$]-putrescine, [$^{2}H_{6}$]-salicylic acid and [$^{13}C_{6}$]-glucose), were dissolved in water. The final concentration of each reference compound in the solvent mixture was 15 ng/ µL. The extraction time was investigated (0.5, 1 and 1.5 minutes) and during each extraction the mill was also stopped twice and, according to the design, solvents were either added or not added to the tubes. For a fraction of the samples the tubes were heated at 60°C for 15 min in pre-heated metal containers.

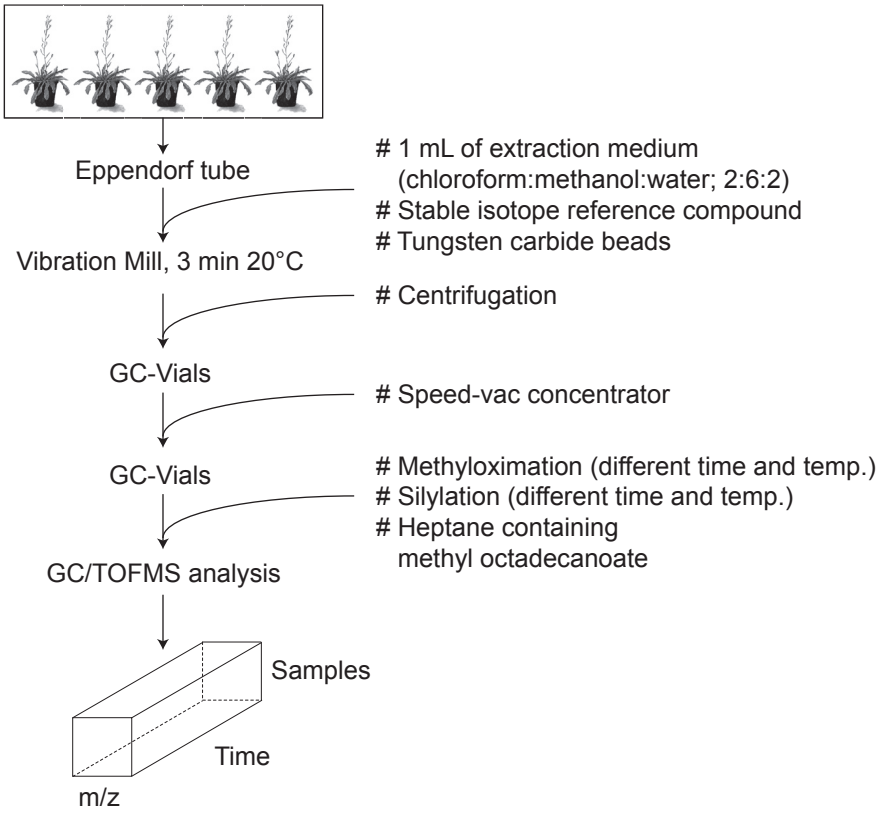


Figure 9. Flow chart of plant metabolomic analysis including plant extraction, oximation, and silylation before GC/MS analysis.

Nearly half of the extracts were then stored at -80°C for 24 hours. After extraction the samples were centrifuged in an Eppendorf centrifuge (model 5417C) for 10 min at 14000 rpm and 180 µL of each supernatant was transferred to a GC vial and evaporated to dryness in a Speed-Vac concentrator. Thirty µL of 15, 20 or 25 mg/ml methoxyamine in pyridine were added, and the mixtures were then vigorously stirred for 15 minutes. The aldehyde and keto-groups were converted into oximes either at room temperature for 16 hours or heated at 60°C for 1 hour. Using 15, 22.5 or 30 µL MSTFA followed by vortex-mixing for five minutes the –OH, –SH or –NH groups were thereafter converted to trimethylsilyl (TMS) derivatives. To each vial heptane was added to compensate for differences in the final volumes. All samples were prepared at the same time, analyzed in randomized order and after at least one hour at room temperature the samples were injected into the GC/MS.

GC/MS Metabolomic Analysis

Each derivatized sample (1 µL) was injected splitless by an Agilent 7683 autosampler (Agilent, Atlanta; GA, USA) into a 10 m x 0.18 mm i.d. fused silica capillary column (0.18 µm DB 5-MS stationary phase, J&W Scientific, Folsom, CA, USA) in an Agilent 6890 gas chromatograph. The injector temperature was 270°C, the purge

flow rate was 20 mL/min and the purge valve was turned on after 60 s. The column temperature was kept at 70°C for 2 minutes followed by ramping at 40°C/min to 320°C, which was held for 1 min. The helium gas flow rate through the column was 1 mL/min. The column outlet was connected to the ion source of a Pegasus III time-of-flight mass spectrometer (TOFMS, Leco Corp., St Joseph, MI, USA). The transfer line and ion source temperatures were 250°C and 200°C, respectively. Ions were produced by a 70 eV electron beam at an ionization current of 2.0 mA. Mass spectra were recorded in the mass range 50 to 800 m/z at a spectra accumulation speed of 30 spectra/s. For the extraction design and reproducibility test reported in Paper I only 80 to 800 amu spectra were recorded. The acceleration voltage was turned off and no mass spectra were detected during the first 170 s. The detector voltage was 1500 V.

Table 4. Factor settings for the $2^{(7-1)}$ fractional factorial design used for optimising the extraction and derivatization. E=extraction; D=derivatization. *Chloroform alone = use of chloroform alone during the first extraction period; **Two phases = use of water and chloroform together, with no other solvent, during the extraction. ***Balance, heptane.

Nr	Factor	Abbreviation	Settings
1	E_Chloroform_Alone*	E_CA	Alone or Not alone
2	E_Phase	E_Ph	One phase or Two phases**
3	E_Heat	E_He	20 or 60°C
4	E_Freeze	E_Fr	0 h or 24 h
5	D_Oxime_Amount	D_OA	14 or 25 mg/ml
6	D_Time_Temp	D_TT	60°C for 1h or 20°C for 16h
7	D_MSTFA	D_MS	50 or 100 % ***
	Replicate points		Chloroform alone, one phase, 37°C, 0 h in freezer, 19 mg ml^{-1}, 60°C for 1h and 75 % MSTFA***

Standard Mix Example

Three standard compounds ([$^2H_4$]-succinic acid, succinic acid and norleucine) that co-elute in the GC/MS analysis after derivatization (Figure 5) were chosen to investigate the processing strategies developed and described in Papers II and III. The standards were dissolved in water and added in varying amounts and proportions to twenty GC-vials. The solvent was evaporated in a Speed-Vac concentrator. The amounts of labelled and unlabelled succinic acid injected on the column were selected according to a 4^2 design (samples 1-16), plus four centre points (samples 17-20), with total levels of 5, 10, 15 and 20 ng. The amount of norleucine added was 15 ng in each case. No succinic acid was added to sample 20. The samples were analyzed by GC/TOFMS after methyloximation and trimethylsilylation. The data were aligned and one window was set around the three co-eluting peaks according to Paper II.

GA Biosynthesis and Signalling Mutants

Sterile PCR tubes (0.2 mL Thermo-Strips, Cat # AB-0266, Abgene, Epsom, UK), filled with 1 x MS (Murashige & Skoog, 1962) and 0.8% agar with a pH of 5.6, were cut at the bottom and placed in sterile boxes. The boxes were black so only the top of the tubes could be exposed to light. The bottoms of the tubes were in contact with

sterile 0.5 x MS media. Sterilized seeds were treated with 10^{-5}M GA_4 in darkness for 72 h at 4°C. After washing with sterile H_2O the seeds were transferred to the tubes in 0.1% Agarose (one seed per tube) Approximately 450 seeds per genotype were sown for five different treatments. For each of eight genotypes, leaves and stem material from three plants grown under long day conditions for 18 days (3-4 leaf stage) were pooled in eppendorf tubes and immediately frozen in liquid nitrogen. The eight genotypes were: *ga1-3, gai, gai-t6 ga1-3, gai-t6 rga24, gai-t6 rga24 ga1-3, gai-t6 rga24 sly1-10, rga24 ga1-3*, *sly1-10* and WT. For the remaining plants the medium was changed to fresh 0.5 x MS medium. Half of the remaining plants were treated with 10^{-5}M GA_4. After 24 and 48 hours both the GA-induced and non-induced plants were harvested. The number of replicates analyzed varied from three to seven pooled samples. Plants were extracted, derivatized and analyzed using GC/MS according to the protocols described below. The amount of plant material varied between 10.7 and 260.3 mg.

Extraction and Derivatization of Plant Samples

The extraction procedure used in both the metabolomics and IAA analyses, and the following derivatization (Paper IV) was similar to that previously described in the chapter on derivatization and extraction design. All samples were randomized and divided into three batches. Samples from each batch were extracted and analysed in a three-day period according to the same procedure. [$^{13}C_6$] IAA was added as an internal standard to the mixture of the three solvents for the IAA analyses. The material was extracted for 3 min then centrifuged. To compensate for differences in the weight of the samples, the volumes of the plant extracts used for the metabolomic analyses were adjusted accordingly. The volume of the supernatant transferred to each GC vial corresponded to 4 mg of plant material. To each vial stable isotope reference compounds for metabolomic analysis were added and evaporated to dryness. The metabolomic samples were derivatized using 30 µL of methoxyamine hydrochloride (15 mg/mL) in pyridine then incubated at 70°C for one hour, followed by 16 h of derivatization at room temperature. The samples were trimethylsilylated for 1 h at room temperature by adding 30 µL of MSTFA with 1% TMCS. After silylation, 30 µL of heptane containing 45 ng/µL methyl octadecanoate was added. The samples (in total 254) were analyzed in a three-day period by GC/TOFMS together with blank samples and alkanes (C_{12}-C_{40} series) series. For the IAA measurements (Paper IV) 500 µL of plant extract was dried and then dissolved in 20 µL methanol followed by 0.5 mL of 0.05 M phosphate buffer, pH 7.0. The pH was adjusted to 2.7 using 1 M HCl. The solution was applied to a 500 mg C_8 Bond Elut SPE column (Varian, Harbor City, CA, USA, conditioned with 2 mL of methanol and 2 mL of 1% acetic acid) at a flow rate of less than 1 mL/min. The columns were washed with 2 mL of 10% MeOH in 1% HAc and the samples were eluted with 2 mL MeOH. After evaporation to dryness the samples were methylated using 200 µL 2-propanol, 1000 µL dichloromethane and 5 µL (trimethylsilyl)diazo methane (2M in hexane). The samples were left at room temperature for 30 minutes, 5 µL of acetic acid was added to each of them, and they were dried in a speedvac concentrator. The samples were transferred in methanol to GC vials and the solvents were evaporated. Each sample was trimethylsilylated by adding 10 µL pyridine followed by 10µL BSTFA with 1% TMCS and heated to 70°C for 30 minutes. 50 or

150 µL of heptane was added after evaporation to dryness, depending on the amount of [$^{13}C_6$]IAA added to each sample. The samples were analyzed by the GC/MS-selected reaction monitoring (SRM) technique previously described by Edlund *et al.* (1995) using a JEOL MStation mass spectrometer (JEOL, Tokyo, Japan).

Experimental Design and Multivariate Data Analysis

All manual integrations were performed using ChromaTOF 1.00 software (Leco Corp., St Joseph, MI, USA). In the studies described in Papers I, II and IV automatic peak detection and mass spectrum deconvolution were performed with a peak width set to 2.0 s and peaks with lower signal-to-noise (S/N) values than 10 were rejected. The software calculates the S/N based on the masses it chooses for quantification. Selected unique quantification masses for each metabolite were used for peak area determinations. Mass spectra of all detected compounds were compared with spectra in NIST library 2.0 (as of January 31, 2001), and in-house and publicly available databases (Schauer *et al.*, 2005). All experimental designs were generated and evaluated using MODDE (Umetrics, Umeå, Sweden). All multivariate investigations for both PCA projections and PLS calibrations were performed using SIMCA-P software (Umetrics, Umeå, Sweden). For all PLS models, the variables were both mean centred and scaled to unit variance except in Paper IV where only scaling to unit variance was performed. For Paper I factors that did not improve the experimental model according to cross-validation were removed before interpretation.

All OLS and PLS calculations were performed using 95% confidence intervals for the metabolomic analyses and 99% confidence intervals for the IAA investigations. In the studies described in Papers II, III and IV non-processed MS-files were exported from the ChromaTOF software in CSV format to MATLAB™ software 6.5 (Mathworks, Natick, MA, USA), in which all data pre-treatment procedures, such as base-line correction, chromatogram alignment, data compression and curve resolution were performed using "in house" custom scripts. In Paper IV non-normalized time weight vectors were calculated using MATLAB™. The variation in the metabolite matrix presented in Paper IV to predict each genotype and each design factor was analysed according O-PLS. The numbers of significant O-PLS components for the calibration models were estimated according to full cross-validation.

Results and Discussion

Optimization of the GC/MS Plant Metabolite Protocols

GC/TOFMS systems enable mass spectra to be accumulated rapidly (Veriotti & Sacks, 2001), making them highly suitable for the analysis of complex mixtures (Weckwerth *et al.*, 2004), such as metabolomic samples from *Arabidopsis* (Figure 3). The advantage of rapid analytical cycles, around 15 minutes per sample, is that they allow high through-puts, in our case 90 samples per 24 h. The overall result of the analysis will be dependent on both the instrumental settings (O'Hagan *et al.*, 2005) and the procedures applied during extraction and derivatization. DOE was used together with multivariate analysis to investigate how different parameters (choice of extraction solvents, derivatization reagents and physical conditions) affect the extraction and derivatization conditions of plant metabolites. This was done in three stages: (1) screening for factors affecting oximation and silylation using 12 metabolites that commonly occur in plant tissues; (2) investigation of how different extraction and derivatization conditions affect the detection of metabolites from *Arabidopsis*; and (3) investigation of the reproducibility of suggested extraction and derivatization methods.

Screening of Factors Affecting Derivatization

In the study described in Paper I a number of chemicals were selected to cover different classes of compounds and a wide chromatographic retention span in the GC. Factors that have a strong effect on methyloximation and silylation were chosen for investigation. A D-optimal design was generated for this series of experiments since we wanted to investigate both mixture factors (i.e., heptane and MSTFA) and process ("regular") factors (i.e., temperature and time) in the same design (Table 3). The advantage of performing a separate derivatization screening was that compounds that may decompose during derivatization and artefacts caused by derivatization could be monitored. Fourteen PLS models with two components were calculated and validated using cross-validation for the twelve standard compounds, the number of peaks detected by the ChromaTOF software for each sample and the peak area ratios for fructose and sucrose. The number of peaks estimated by the ChromaTOF software was valid according to the cross-validation. About 20 measurable peaks expected from the 12 derivatized standards, and on average 238 peaks in total, were detected, indicating that this kind of analysis generates a high number of artifacts (Little, 1999). This was highly dependent on the amount of MSTFA, the temperature and the amount of pyridine present during oximation. On the other hand, the amounts of MSTFA and pyridine also had a positive effect on the response for many of the compounds. During the oximation, temperature and/or time played important roles in the derivatization of glucosamine and alpha-ketoglutaric acid, for complete derivatization of which a long time and/or high temperature was needed. Other compounds, such as sucrose, were also clearly affected by the time and temperature, but it seems that both high temperature and long durations of oximation are required for the decomposition of sucrose to fructose and glucose. In contrast, the temperature during the silylation did not have any dramatic effect on the tested compounds.

Screening of Factors Affecting Extraction

Traditional analytical protocols focus on analysing limited numbers of specific compounds. When the goal instead is to analyze as many metabolites as possible from a biological tissue, it is a considerably more complex task. It is practically impossible to ensure high accuracy and precision for all metabolites since it is not possible to add stable isotope-labelled internal standards for all of the detected compounds. Furthermore, it is also impossible in practice to extract all metabolites efficiently since plant tissues (and other biological materials) contain metabolites that differ widely in their chemical nature and amount. Therefore, the goal must be to develop an extraction method with as high extraction efficiency and reproducibility as possible for as many classes of compounds as possible. The extraction buffer selected will dramatically affect the type and number of metabolites extracted from plant tissues. Therefore, the efficacy of various solvents, including MeOH, EtOH, acetonitrile, chloroform and hexane, was evaluated in a pilot study. Seven to sixteen percent fewer peaks were extracted for a GC/MS analysis using water or methanol alone as extraction solvents compared with a chloroform:MeOH:water mixture. Chloroform had a positive effect on the extraction of lipophilic compounds, such as fatty acids. In the study described in Paper I chloroform:MeOH:water in the ratio 2:6:2 was chosen because it allowed the solvents to be used in a single mixture and avoided solvent partitioning. Sixty-six metabolites were selected to measure the effects of different extraction and derivatization protocols in an extraction design (Table 4). All labelled internal standards could be analyzed, and more or less all of the compounds could be detected in the 68 samples. Variations in chromatography or amounts of plant material were minimized by dividing the response by the response for methyl octadecanoate and the respective tissue weight. Sixty-eight OLS models were calculated, one for each peak area response, one for the number of peaks and one for the degradation of $^{13}C_{12}$-sucrose to $^{13}C_{6}$-fructose. The numbers of the 66 metabolites that were significantly influenced by each of the 28 design factors, which can be seen in Table 5. Extraction with chloroform alone and use of a two-phase system has a positive effect on the metabolite extraction efficiency. High temperature during the extraction has a positive effect. The number of peaks detected by the software is highly dependent on the amount of MSTFA added, as shown in the derivatization design. The amount of MSTFA added was also important for the peak response and for 66 of the chosen metabolites these two variables were positively correlated (Table 5). The oximation method also had a strong effect on the results of the derivatization. The concentration of methoxime influenced the response, and the response for many peaks was favoured by high temperature during oximation. When developing methods for extracting metabolites it is important to minimize their chemical or biological degradation. As an indicator of biological (or other) degradation, the hydroxylation of $^{13}C_{12}$-sucrose to labelled glucose and fructose was measured. The main factor causing degradation was the order in which the different extraction solvents were added. Adding chloroform alone, and using a two-phase system (i.e. chloroform and H_2O together, prior to the addition of MeOH) had a dramatic effect on degradation. Although the oximation method also affects the degradation of sucrose, the main factor influencing the degradation is the extraction method.

Table 5. Factors that significantly (p=5%) affected the 66 endogenous metabolites in the extraction and derivatization design. *Amount of MSTFA (%)

	Positive	Negative
E_Ch loroform (E_CA)	3.0% (Not alone)	30.3% (Alone)
E_Phase (E_Ph)	24.2% (Two phases)	6.1% (One phase)
E_Heat (E_He)	53.0% (60 °C)	7.6% (20 °C)
E_Freeze (E_Fr)	0.0% (24 h)	22.7% (0 h)
D_Oxime_Amount (D_OA)	12.1% (25 mg/mL)	27.3% (14 mg/mL)
D_T ime_Temp (D_TT)	7.6% (20 °C for 16 h)	45.5% (60 °C for 1h)
D_MSTFA (D_MS)	45.5% (100%[a])	12.1% (50%[a])
E_CA (Not alone)*E_Ph (Two phases)	0.0%	9.1%
E_CA (Not alone)*E_He (60°C)	1.5%	21.2%
E_CA (Not alone)*E_Fr (24h)	28.8%	3.0%
E_CA (Not alone)*D_OA(25 mg/mL)	0.0%	3.0%
E_CA (Not alone)*D_TT (20°C for 16h)	21.2%	0.0%
E_CA (Not alone)*D_MS (100%)	0.0%	0.0%
E_Ph (Two phases)*E_He (60°C)	1.5%	7.6%
E_Ph (Two phases)*E_Fr (24h)	7.6%	1.5%
E_Ph (Two phases)*D_OA (25 mg/mL)	9.1%	0.0%
E_Ph (Two phases)*D_TT (20°C for 16h)	3.0%	7.6%
E_Ph (Two phases)*D_MS (100%)	6.1%	6.1%
E_He (60°C)*E_Fr (24h)	4.5%	1.5%
E_He (60°C)*D_OA (25 mg/mL)	1.5%	9.1%
E_He (60°C)*D_TT (20°C for 16h)	16.7%	3.0%
E_He (60°C)*D_MS (100%)	7.6%	4.5%
E_Fr (24h)*D_OA (25 mg/mL)	3.0%	21.2%
E_Fr (24h)*D_TT (20°C for 16h)	1.5%	0.0%
E_Fr (24h)*D_MS (100%)	3.0%	3.0%
D_OA (25 mg/mL)*D_TT (20°C for 16h)	3.0%	4.5%
D_OA (25 mg/mL)*D_MS (100%)	10.6%	15.2%
D_TT (20 °C for 16 h)*D_MS (100%)	9.1%	4.5%

To develop a satisfactory procedure for both extraction and derivatization of a large number of metabolites compromises have to be made, in which all of the factors mentioned above must be taken into account. In the selected protocol chloroform was used at the start of the extraction, since it has a positive effect on extraction efficiency, and causes relatively little degradation of sucrose. The vials were heated to 60°C after extraction. In order to dissolve certain metabolites that are difficult to dissolve and to oximate completely some that are resistant to this process, the oximation was performed at 60°C for one hour followed by 17 hours at room temperature. The amounts of methoxime and MSTFA added were 20 mg/ mL and 30 μL, respectively. The extraction protocol inevitably involves a compromise between efficiency and speed. The solvent mixture chosen in the present protocol extracts both hydrophilic and lipophilic compounds. According to our study, 80-100% MeOH will not efficiently extract polar compounds. There is a risk of transmethylation of sugar esters using MeOH, but we still included MeOH since we found it to be a

more efficient extraction solvent than EtOH or ACN. In metabolomic analyses the extracts are often divided into polar and lipophilic fractions by solvent partitioning (Broeckling *et al.*, 2005; Desbrosses, Kopka & Udvardi, 2005). In our extraction procedure no solvent partitioning step after extraction was used, in order to minimize the time required and to facilitate automation. Increasing the sample preparation time and the number of injections of each sample clearly decreases the scope for high-throughput analysis, which is detrimental for metabolomic studies since they generally involve the analysis of large numbers of samples. To be able to generate accurate and precise GC/MS data, labelled stable isotope internal standards (Fiehn *et al.*, 2000a; Broeckling *et al.*, 2005) are added during analysis. We chose to add 11 internal standards representing different classes of compounds, e.g. amines, amino acids, fatty acids, sterols, mono- and di-saccharides. In this way, accurate levels for quite large numbers of compounds can be determined (Fiehn *et al.*, 2000a) rather than only relative levels (Roessner *et al.*, 2000). The selected protocol was used to investigate the reproducibility of the extraction. With this extraction procedure it was possible to detect and quantify the corresponding endogenous metabolites for seven of the labelled reference compounds and the remaining 59 metabolites were quantified using [$^{13}C_3$]-myristic acid as an internal standard. After correction for the difference in responses of the internal standards, the mean errors were 8.2% (6.9-9.7%) and 13.8% (5.5-33.4%) for metabolites with and without a specific internal standard, respectively; similar to errors reported using other metabolomic approaches (Roessner *et al.*, 2000). There will always be quantification problems in global analyses of metabolites since it is impossible to include a labelled standard corresponding to every one of them. Due to the normalization problem, correlations between the normalized metabolites and the time between silylation and injection into the GC/MS were investigated. A PLS-model was calculated between the metabolite matrix (**X**) and the time vector (**y**). According to the cross-validation, the first four PLS-components were significant (R^2X=0.58, R^2Y=0.98 and Q^2Y=0.75). The first weight vector (**w**) from the PLS-model was used to find significant correlations to the injection order by Jack-knifing (Efron, 1986; Martens & Martens, 2000), and eighteen of the 66 metabolites were found to be significantly correlated to the injection order. Of the seven metabolites normalized using corresponding labelled reference compounds none showed significant variation correlated to the injection order. This addresses the problem of not using unique internal standards for normalization of the integrated peaks, which can be controlled by randomizing the run-order. The problems associated with the correlation between run-order and response can then be minimized, thereby reducing the overall method error. An advantage of using more than one internal standard is that errors during analysis and derivatization can be more easily detected for different classes of compounds. Thus, the quality of the analysis can be checked simply by comparing peak areas for the different internal standards in the various samples. By increasing the number of internal standards it would be possible to further improve the reproducibility. Although the errors could also be decreased by improving the chromatographic system and peak area integration, most errors originate from weighing, pipetting and, to some extent, irreproducible homogenization/extraction. However, it is difficult to improve these parameters, as ease and rapidity are key concerns in metabolomics. For example, a more efficient extraction would probably demand longer, repeated extractions.

In the present investigation, we have shown that it was possible to generate a

reliable protocol for metabolomic analysis of *Arabidopsis* using this strategy with a relatively limited number of experiments. Without using DOE it would have been very difficult, without using many more experiments, to obtain an overview of the problems associated with optimizing the extraction and/or derivatization protocol for metabolomic analysis.

The method presented here is rapid, and involves steps that can be quite easily automated. The oximation step is quite long (overnight at room temperature), but it could be shortened, while maintaining high oximated product yields for most of the compound classes. As every method essentially represents a compromise between many variables, a fully automated extraction and derivatization method based on the principles of the present method could be developed. Furthermore, we have shown that the strategy can be very efficient for optimizing extraction conditions for metabolomic analysis.

Hierarchical Methods for Processing GC/MS Data

The rapid mass spectra accumulation of the GC/TOFMS systems (30 spectra/s, corresponding to 20-40 data points per peak) makes it possible to analyze plant samples in a high thought-put manner. The mass spectra are homogenous over the peak profile which simplifies the deconvolution and data interpretation. However, it is not straightforward to resolve unique peak information in complex chromatograms containing hundreds of overlapping peaks rapidly (Figure 3). The traditional way of comparing GC/MS data sets is to resolve chromatographic overlaps in the MS-files, then calculate the relative amounts of each compound and finally subject the data to statistical modelling. Resolving chromatographic peaks has been a time-consuming process to date, and evaluation of the enormous amounts of data generated is therefore a bottleneck in the "metabolomics" era, especially for high through-put analyses. In Papers II and III we present two semi-automated strategies that enable rapid comparison of non-processed MS-data files from metabolomic analyses. The primary goal for us is to identify the differences between samples, not generate lists of peaks. The first method, presented in Paper II, Method 1, is a hierarchical bilinear compression method, compressing the information from small retention windows and generating approximations of the compounds' intensity and their corresponding mass profiles. The second method, presented in Paper III, Method 2, applies Hierarchical Multivariate Curve Resolution (H-MCR) to the same retention windows to generate intensity values for peaks and their corresponding mass spectra. Both methods share the same starting pre-treatments of the sample; smoothing, background reduction, alignment and time-window setting. MS-files from a metabolomics project were exported in NetCDF or CSV-format from ChromaTOF software to MATLAB™ software for further processing. For each sample each m/z-channel is smoothed with moving averages (each of seven time-points) followed by background reduction by subtracting the minimum value of each m/z channel from all other values in the same m/z channel. This is important when dividing the chromatograms into time-windows since it makes the start and end points of peaks or peak clusters easier to detect. All chromatograms were aligned so that all samples had maximum covariance between the Total Ion Currents (TIC) according to Malmquist & Danielsson (1994). The alignment compensates for small differences in retention between different injections, which will always occur. Even after alignment the drifts may not remain

constant throughout the course of each chromatographic analysis, so each file is divided into small time-windows. A time window is defined as a short retention span that starts and ends in a region of the chromatogram that does not contain any compound (low intensity points). The step is done manually and simultaneously for all samples. This is the only manual step in both Methods 1 and 2; all other steps are fully computer automated. The following steps used to compare the GC/MS samples in the two methods are summarized in Figure 10.

Method 1

The data in each time-window can be seen as a data cube of size N*K*L, where N = the number of samples, K = the number of time points (scans) and L = the number of m/z channels. In each time-window there can be several compounds (components). The two strategies use different methods to estimate the components' intensity (C) and their corresponding mass profiles (S). The first method summarizes each window into a mass spectrum to generate a sample matrix (N*L). It is not problematic to compare the summarized mass spectrum between each sample since the multivariate modelling will still recognize the differences between samples. To extract information from each time-window AR is used. To do this the chemical rank (number of components) has to be known for the window matrices (i.e. the summarized mass spectra of all samples in the respective windows; Figure 10). This estimated rank is the number of eigenvalues above the noise level calculated using PCA (Liang & Kvalheim, 2001). As a starting estimate of S, a random number is used. During the iteration according to equations 8 and 9, both S and C are corrected to have non-negative constraints. This means that the concentration and the mass spectra can never include negative values. Negative values in the spectrum profile are set to zero, and negative concentration values are set to the lowest of all positive concentrations. AR alters the alignment until convergence according to Karjalainen (1989). The chemical rank is reduced by one if the correlation between spectral profiles is greater than 95% or if the AR-algorithm does not find a solution.

Method 2

For the second method the cubes are unfolded to a matrix built up by placing each sample on top of each other to generate a **X** matrix of size (N*K) * L (Fig. 10). This data matrix is then resolved using AR. As starting estimates of S we have used the purest mass channels calculated according to the SIMPLISMA algorithm which gives a more rapid convergence than using only random numbers as start estimates. The estimate of C will be of size (N*K) * the number of resolved components (R; i.e. the number of components with a common mass spectrum). In contrast to decomposition of one sample using AR, where only one peak is estimated for each profile in C, in this case the profiles in C will consist of a number of peaks (Figure 10). The same, unimodal and positive, constraints were used during AR alternating according to equations 8 and 9 as in Method 1. In addition to those constraints, each resolved component must be unimodal for the chromatographic profile for each sample. The resolved component must also elute around the same retention time in every sample. The H-MCR procedure starts with the assumption of a rank of 1 and estimates S and C. The number of components is continuously increased by one until three subsequent solutions are rejected as not valid (the last valid solution is

used). To accept peaks we have a time criterion between peaks of ± 0.5 s from the median retention time for that component. There is no limitation on the number of resolved components that can be found in each time-window. To be able to resolve components using Method 2 a number of criteria have to be fulfilled: the chromatographic profiles must differ or the component ratio between the samples has to differ in the mass spectral dimension. This is advantageous when resolving pure profiles in comparison with method 1, where only the second criterion can be applied due the summarization of the chromatographic direction.

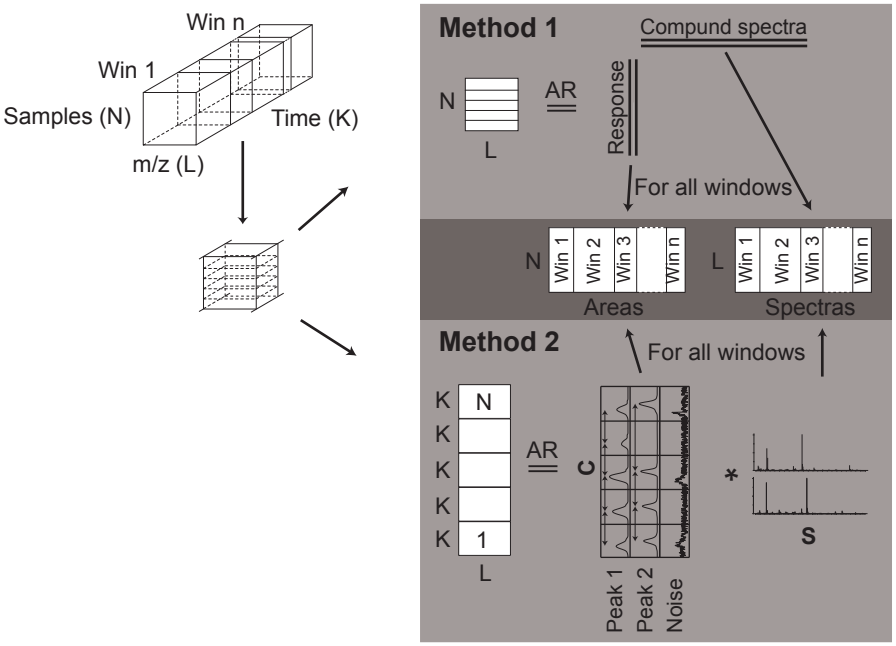


Figure 10. Summary of the metabolomic analysis concept using hierarchical methods for automatically resolving GC/TOFMS data (Paper II, Method 1 and Paper III, Method 2). The data can be used for further multivariate analysis and the mass spectra that explain the differences between samples can be used for identification of metabolites and further biological interpretation. C corresponds to Chromographic profiles, P to peaks, and S to spectral profiles.

Mixture of Standards

The two different strategies are exemplified and validated with two sets of samples: a mixture of standard compounds and *Arabidopsis* samples. The first example is based on results from 20 samples consisting of methoxymated and trimethylsilylated [$^2H_4$]-succinic acid, succinic acid and norleucine analysed by GC/TOFMS. The resolved mass spectra were compared with standard mass spectra, and by calculating the correlation between estimated amounts of the metabolites and the injected amounts. One time-window was selected near the three peaks. For Method 1 the chemical rank was automatically estimated as two and the number of estimated compound using Method 2 was five. One mass profile calculated with the first method had a match calculated by NIST MS Search 2.0 of 891 with norleucine-TMS and the second

corresponded to a linear combination of mass spectra of both [$^2H_4$]-succinic acid and succinic acid. By summarizing the added concentration of both the endogenous and labelled succinic acid a correlation of 97.7% (r^2) was obtained with manually integrated areas for succinic acid. The five deconvoluted chromographic profiles, c_1-c_5, for the 20 samples obtained with the second method can be viewed in Figure 11.

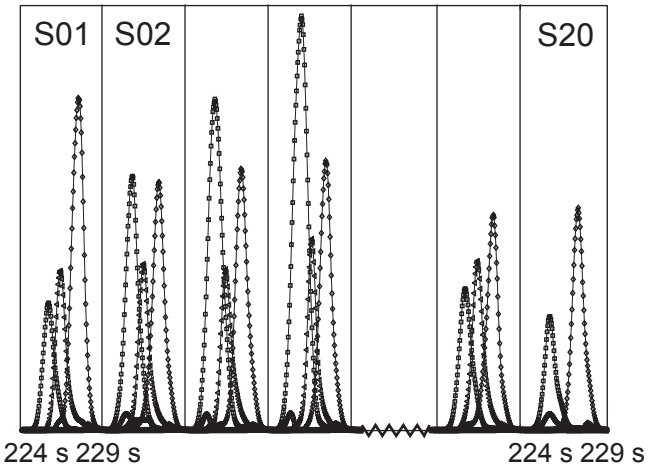


Figure 11. Five resolved chromatographic profiles obtained using the H-MCR method for 20 samples. The results correspond to the analysis of mixtures of three samples containing the standard mixture. Circles correspond to c_1 (225.21 sec, some similarity with [$^2H_4$]-succinic acid-TMS), squares to c_2 (225.87 sec, [$^2H_4$]-succinic acid-TMS), triangles to c_3 (226.57 sec, succinic acid-TMS), diamonds to c_4 (227.34 sec, norleucine,N,O-TMS) and six-pointed stars to c_5 (227.84 sec, some similarity with norleucine,N,O-TMS).

Three of the profiles (c_2, c_3 and c_4) have on average about 50 to 100 times higher peak areas than the remaining c_1 and c_5. The mass profile for component 2 has a match of 943 to [$^2H_4$]-succinic acid-TMS, the third a 923 match with succinic acid-TMS and the forth component a 935 match with the mass spectrum of norleucine N,O-TMS. Component 1 shows some similarity to [$^2H_4$]-succinic acid-TMS and the last component shows some likeness to norleucine (match < 650). [$^2H_4$]-succinic acid and succinic acid, levels of which were changed according to the design, both had r^2 values > 99% for components 2 and 3. For the first method a problem is to estimate the right chemical rank. In this case when the rank was estimated to be two one of the components represented a mix of the different analytes. For the second method a much better estimation of the contributory components to the C and S profiles was achieved. One problem with the second method is that it can introduce too many components; a common problem in deconvolution.

The Arabidopsis *Mutant Test Case*

To further test the two GC/MS processing methods metabolites were extracted from leaves of seven *Arabidopsis thaliana* GA biosynthesis and signalling mutants, and analyzed by GC/TOFMS after derivatization according to the method described in Paper I (Paper IV). The *rga24 ga1-3* plants were difficult to grow and were

excluded from both the IAA and metabolomic analyses. The plants were subjected to an experimental design for two factors, GA_4 treatment and time after treatment. The plants can be dived into two groups, those that were subjected to 10^{-5} M GA_4 treatment and those that were not. The GA_4-treated plants were sampled after 24 and 48 hours, and the control plants at 0, 24 and 48 hours. Due to limitations in computer memory it was not possible to assign large enough retention time-windows for large series of samples, so the peak area around sucrose was manually integrated. The GC/MS information for the eight different plants (207 samples in total) was thereby divided into two retention time blocks, the first from 184.1 to 425.3 seconds and the second from 435.8 to 526.2 seconds. The two blocks were manually divided from base peak chromatograms into 90 and 30 retention time windows, respectively. Using the H-MCR method (Method 2) 386 components in the 120 time windows were resolved, and 624 components were obtained by Method 1. Results from the two methods for control samples and the samples treated with GA_4 were compared by calibrating the two **X** matrixes for the remaining 167 samples (excluding time zero control samples) with a number of **y** variables using O-PLS (Trygg & Wold, 2002; Table 6). The **y** responses consisted of the three design factors from the plant experiment, run order for GC vials and eight "dummy" vectors representing the different genotypes. Each dummy vector consisted of zero values in each row except for samples belonging to the genotype. Models obtained using the two methods show similar ability to predict the different responses (Q^2Y). The calibration model based on the metabolite information generated from Method 1 uses on average a smaller fraction of the variation in the **X** matrix to predict the **y** response (R^2X_{corr}).

Table 6. The metabolic information for the control samples and samples treated with GA_4 at 24 and 48 hours extracted by the two methods, M1 and M2, were compared using O-PLS. **y** responses consisted of the three design factors from the plant experiment, run order for GC tubes and eight "dummy" vectors representing the different genotypes. R^2X_{corr}, representing the variation in **X**, is used to predict variation in **y**, and R^2X_{yo}, the variation in **X** due to structured noise, those variations that are independent of **y**. Q^2Y is the explained variation in **y** and indicates the predictive ability of the model. Components consist of one O-PLS component and an additional orthogonal component. The synergistic effect GA treatment and time was not valid according to cross validation.

	R^2X_{corr} (%)		R^2X_{yo} (%)		Q^2Y (%)		Comp	
	M1	M2	M1	M2	M1	M2	M1	M2
GA treatment	1.8	2.9	25	44	57	60	1+5	1+6
Time	2.2	4.0	21	30	50	48	1+4	1+3
Int. GA treatment and time	-	-	-	-	-	-	-	-
Run order for GC vials	8.0	11	15	19	96	96	1+3	1+2
ga1-3	1.1	2.0	28	31	31	30	1+6	1+3
gai	1.7	3.0	28	30	63	56	1+7	1+3
gai-t6 ga1-3	1.7	2.1	19	45	50	61	1+3	1+6
gai-t6 rga24	2.6	3.8	27	46	73	81	1+6	1+7
gai-t6 rga24 ga1-3	1.1	1.4	25	42	52	63	1+5	1+5
gai-t6 rga 24 sly1-10	1.7	2.5	25	45	61	71	1+5	1+6
sly1-10	1.6	2.7	27	44	57	71	1+6	1+6
WT	1.3	1.7	24	41	55	67	1+5	1+5

The variation introduced by run order of the GC vials, derivatization effects and instrumental drifts amount to roughly ten percent of the total variation for both

methods. For all O-PLS models the two methods give similar interpretation of the first score vector. The projections showed similar separation between design factors and genotypes.

Four local O-PLS models were obtained for each mutant genotype and WT for data produced by the second method. The **y** response corresponded to the run order for GC tubes, GA treatment, time and the interaction term between GA treatment and time. On average 10% of the variation was used in the matrixes to predict GA treatment, 10% to predict the time and 8% to predict the synergistic effect. The remaining variation in the sub matrix is uncontrolled variation, introduced by other sources of variation, associated with biological variation and other analytical variations.

To investigate differences in the chromatographic and spectral profiles obtained in GC/MS analyses from real biological samples generated using the two processing methods, results for 12 internal standards were compared with those generated by manual integration using the ChromaTOF software. Almost all time-windows covering each internal standard have overlapping peaks. For all ^{13}C- and ^{2}H-labelled internal standards the retention time on the GC column was almost the same as that of the corresponding endogenous compound, with slightly earlier retention times for the deuterated internal standards. Figure 12 shows the six resolved chromatographic profiles obtained using Method 2 for the retention window for silylated [$^{13}C_5$, ^{15}N]-glutamic acid.

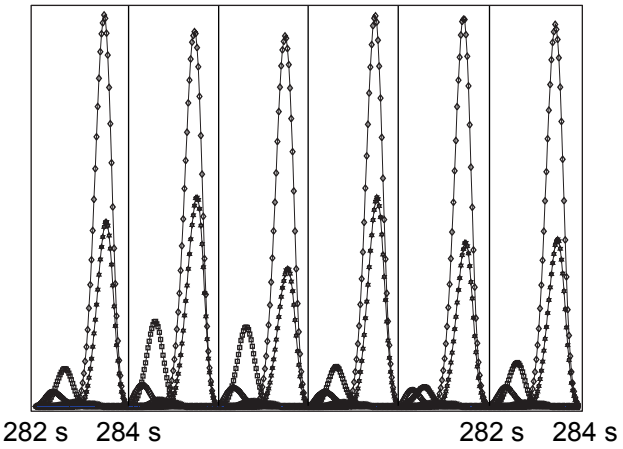


Figure 12. Deconvolution of six chromatographic profiles obtained using the H-MCR method for results from six wt plants harvested after treatment with GA at 48 h. Squares (unknown metabolite eluting at 282.34 s) correspond to the first eluting peak in time window 36. Circles correspond to ornithine (282.68 s), left-pointing triangles to an unknown peak (282.88 s), diamonds to [$^{13}C_5$, ^{15}N]-glutamic acid (283.68 s), six-pointed stars to glutamic acid (283.74 s) and right-pointing triangles (283.84 s) to another unknown peak.

For Method 1 three components were resolved in the time window. Both methods give good estimates of the spectral profiles for both labelled and endogenous derivatized glutamic acid (match >800). The correlations between the manually integrated areas and the estimated responses obtained using the two methods are,

however, different (Table 7). If all response values for three components used to decompose the window using Method 1 are summarized, the correlation with the manually integrated area becomes 44%. The endogenous and labelled compounds have almost identical retention times in the GC/MS system used in the present investigation (Figure 12). Using common deconvolution techniques, e.g. AMDIS or LECO Chromatof™ software, it is not possible to obtain resolved mass spectra for the two compounds. However, with the H-MCR processing strategy described here it is possible to obtain well-resolved mass spectra for both of these compounds, since the ratio between [$^{13}C_5$, ^{15}N]-glutamic acid and glutamic acid differs between samples. The remaining internal standards show similar results as for [$^{13}C_5$, ^{15}N]-glutamic acid in terms of both the correlation between estimated amounts and the injected amounts, and the match of the mass spectra with standard mass spectra. The first method shows on average lower matches to standard mass spectra and mostly lower correlations with manually integrated peaks in comparison with Method 2 (Table 7).

Table 7. Comparison between the two different processing methods, Method 1 (M1) and Method 2 (M2), and their ability to quantify and generate mass spectra for the analysis of a set of *Arabidopsis* mutants defective in GA biosynthesis and signalling, after GA application. The responses for the added internal standard were integrated using a traditional method with Leco Chromatof™ software and compared with the responses obtained by the two methods. The match between the pure spectra of the internal standards and the spectral profiles was performed using NIST MS Search 2.0.

	Rank		r^2 (%)		NIST Match	
Internal standard	M1	M2	M1	M2	M1	M2
[$^{2}H_7$]-cholesterol	1	2	83	87	916	923
[$^{13}C_6$]-glucose	1	5	69	75	640	817
[$^{13}C_4$]- alpha ketoglutarate	3	3	87	86	532	942
[$^{13}C_4$]-hexadecanoic acid	4	3	42	91	929	935
[$^{13}C_5$, ^{15}N]-glutamic acid	3	6	0	93	912	946
[$^{13}C_5$]-L-proline	2	4	77	71	915	934
[$^{13}C_3$]-myristic acid	4	1	90	80	940	944
[$^{2}H_4$]-putrescine	3	5	31	70	820	806
[$^{2}H_6$]-salicylic acid	3	4	56	98	738	935
[$^{2}H_4$]-succinic acid	3	3	91	77	939	941
methyl octadecanoate	2	4	85	83	927	950

There are two types of cases where the two methods do not show high correlations with manually integrated peak areas. One of these cases is when extremely overlapping peaks occur and poor correlation arises from the difficulties in finding unique target masses for quantification with the ChromaTOF™ software. Problems can also occur when very low abundance metabolites are to be quantified. In such cases the poor correlation is usually due to difficulties in finding quantification masses for the ChromaTOF™ software with high enough signal-to-noise (S/N) ratios. This problem was not observed for the internal standards.

An advantage with both methods is that no reference target sample is needed, which makes them less biased than traditional deconvolution methods. The traditional

approach in metabolomics is to use a master sample to find a set of peaks, and thereafter peak-match the other samples. For both Methods 1 and 2 the matching is done for all samples simultaneously, which is preferable for global analyses of metabolites in large series of samples. Both methods generate useful and similar information that can be used for biological interpretation. The first method is focused on compressing information in each time-window and describes the variation. Thus, the time direction is eliminated, so only profiles that differ between samples will be detected. It will always extract profiles that describe the maximum variation in each window. For Method 2 a good estimated solution of the components describing the variation in both the time and mass directions for each time-window will be generated. Applying the criterion that different components should have to have similar retention times gives better estimations of the chemical rank than using PCA. Method 2 also allows mass spectra from two completely overlapping peaks to be resolved, as long as the intensity of the two peaks varies between samples. This gives an improved estimation of spectral and chromatographic profiles. The retention information in combination with mass spectra is of great importance for the identification of compounds (Schauer *et al.*, 2005). However, the retention information obtained with Method 1 (in contrast to method 2, which is restricted to the different time-windows) can also be compared, to some extent, with library information. In the 'omics' era it is usually important to identify the differences between samples, e.g. samples from diseased versus non-diseased, or mutant versus wild-type tissues. It is therefore essential to develop fast data processing methods that do not slow down the whole process from sampling, through extraction, to MS-analysis. Both methods provide fast comparisons and in combination with multivariate techniques/tools make it possible to classify and calibrate samples. One advantage with Method 1 is that the multivariate modelling will be based on smaller data sets, and thus increase the speed of the process. The 207 samples took approximately three days to process using Method 2; almost 10 times slower than Method 1. Method 1 gives a good overview and similar interpretation to Method 2. On the other hand, Method 2 is better for estimating mass spectra profiles and is advantageous when using mass library databases. For quality control of the results, both qualitative and quantitative, they should be referred back to the raw data for both methods. Output data can be analyzed using multivariate statistical tools such as PCA, PLS or O-PLS, and based on the model loadings the peaks (compounds) that differ between sample groups can be identified by comparison with standard library databases.

Effects of GA Biosynthesis and Signaling on IAA Biosynthesis and Metabolite Profiles

In the study described in Paper IV the goal was to study metabolite patterns and IAA levels in *Arabidopsis* mutants lacking GAs and/or parts of the GA-signalling pathways to elucidate interactions between GAs and IAA. This was done by applying bioactive GA_4 (10^{-5}M) to nine genotypes: *ga1-3, gai, gai-t6 ga1-3, gai-t6 rga24, gai-t6 rga24 ga1-3, gai-t6 rga24 sly1-10, rga24 ga1-3, sly1-10* and WT. The experiment was carried out according an experimental design with two design factors: GA treatment (negative and positive) and growth time (24 and 48 hours). Thus, each design factor (time and GA treatment) together with the synergistic effect

of GA treatment and time could be investigated independently of each other. For each genotype four experiments in the design were GA_4-treated samples harvested after 24 and 48 hours, together with corresponding control samples. In addition, control samples collected at time zero were included. All genotypes were extracted, derivatized and analyzed except *rga24 ga1-3* plants, which were difficult to grow and thus were excluded from both the IAA and metabolomic analyses. In total 207 samples were analyzed, 167 of which belonged to the design while the others were time zero control samples.

The goal in the metabolic study was to explore the metabolic effect of GA_4 treatment over time in the different genotypes. The metabolic data were analyzed by GC/TOFMS, the GC/TOFMS data were processed according to H-MCR in Paper III, and in total 386 peaks were resolved. The procedure for data treatment is further described in the *Arabidopsis* mutant test case chapter. Since the H-MCR did not cover the time space of sucrose the corresponding peak area was manually integrated. As described, the variation in the data matrix generated for the plant design and the system, such as the biological differences between the different genotypes, can be calibrated to external information using O-PLS (Table 6). Not all of the variation originates from the experimental design and the genotype differences (Table 6). Other variation is introduced during derivatization and by instrumental drifts, other laboratory analytical variations and biological variation. In order to avoid errors due to chromatographic and analytical variations, all deconvoluted peak areas, and the area for sucrose, were normalized by dividing them by with the score values for the first principle component (PC1) for the internal standards and methyl octadecanoate. Prior to PCA, the areas for the twelve manually integrated peaks were normalized to unit variance and the PC was calculated. The first PC will represent a global average of all internal standards and can be used to compensate for variation during derivatization and GC analysis. Unique OLS models were derived for each peak separately for each mutant and the WT plants. Peaks that were not significantly affected by the GA_4 treatment or by the synergistic effect of treatment and time (in total 103 peaks) for any of the eight different genotypes were excluded from further investigation. By comparing the results for the different genotypes, peaks that were only affected in one genotype were excluded from further multivariate comparisons. In total 202 remained. This was done in order to exclude the peaks corresponding to analytical error, biological variations and normal growth variations.

The only variation in the **X** matrix (integrated peak areas) of interest in this study is the variation related to the time effect after treatment of GA_4. To exclude unrelated structural and non-structural variation the samples corresponding to each genotype were divided into two groups: one that had been treated with GA_4 and one that had not. To each of the two sample groups the corresponding time zero samples were added. For each sub group a PLS model was calculated against the time response and validated according to cross-validation. The **X** matrix was scaled to unit variance and to the **y** response a value of 1 was added. From each of the PLS models non-normalized weight vectors are calculated. This is performed by transposing each peak matrix and multiplying it by the unique time vector according to:

$$(10)\ \mathbf{w}_{\#} = \mathbf{X}^{T}\mathbf{y} / N$$

Where **X** is a sub matrix comprised of 202 integrated peak areas (metabolites), N is the number of samples, and **y** is the unique time vector. The first $\mathbf{w}_{\#}$ will be a

vector corresponding to the time effect for each integrated peak and each separate sub-matrix. Normalization of the weight vector would in this case complicate the comparison of $w_{\#}$ between the different genotypes. For each subgroup $\mathbf{w}_{\#}$ (equation 10) was calculated seven times by Jack-knifing. The calculations of $\mathbf{w}_{\#}$ were performed seven times for each sub-group. For each $\mathbf{w}_{\#}$ calculation, every 7th sample was excluded from the **X** matrix and also for the response vector **y**, in order to obtain a better overview of the stability and the variability of the calculated $\mathbf{w}_{\#}$. Consequently, if the value for one metabolite for one of the genotypes increases over time after GA_4 treatment but has a constant value for the untreated plants, $\mathbf{w}_{\#}$ for the treated plant will have a higher value than for the untreated plant. By comparing $\mathbf{w}_{\#}$ for the different metabolites between the 16 new groups, different effects of the treatment can be seen. The weight vectors for the 202 metabolites, for all sub-groups, were analyzed using unit variance scaled PCA. Clear differences between the induced (n) and not induced (p) genotype could be seen and the effects of GA_4 can be seen in the score plot $\mathbf{t}_2/\mathbf{t}_3$ (Figure 13). In the plot, the data points for the seven samples for each of the 16 subgroups subjected to the Jack-knifing procedure provide a representation of the biological variation.

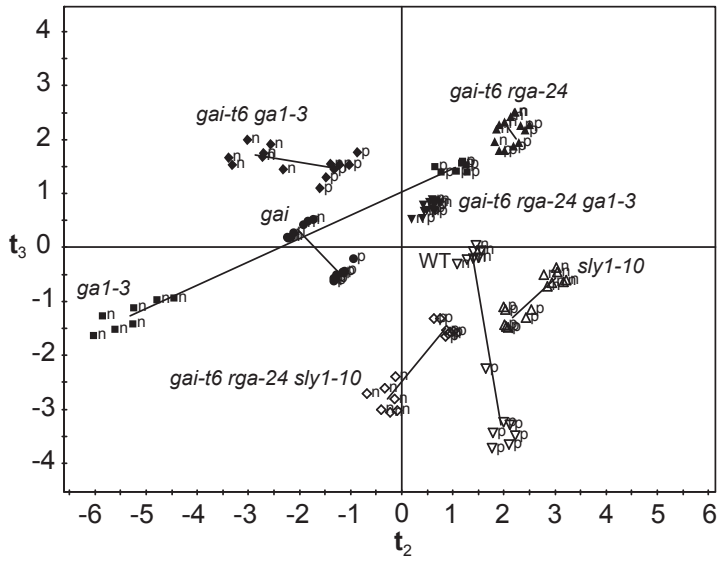
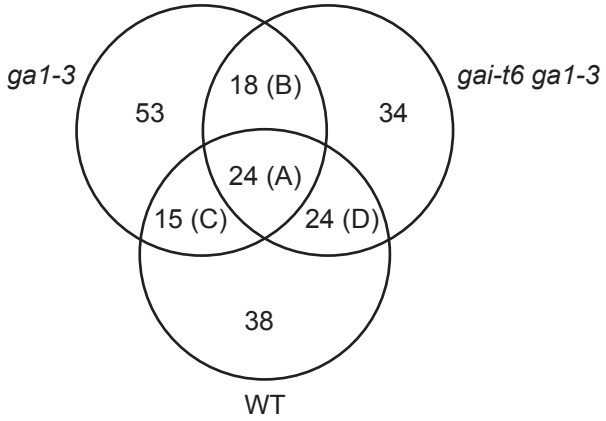


Figure 13. A PCA plot for score vectors **t**2/**t**3 showing differences between the effects of GA treatment on eight different GA mutants in the following 48 hours. The data correspond to the first non-normalized weight vector calculated for 16 PLS models using time as **y** response. Each of the sub matrixes used for the PLS models are represented as seven samples in the plot according to the Jack-knifing routine. Not induced and induced genotypes are represented by n and p, respectively. Squares correspond to the *ga1-3* samples, circles to *gai*, diamonds to *gai-t6 ga1-3*, up-pointing triangles to *gai-t6 rga24*, down-pointing triangles to *gai-t6 rga24 ga1-3*, open diamonds to *gai-t6 rga24 sly1-10,* open up-pointing triangles to *sly1-10* and open down-pointing triangles to WT.

Both the non-GA-treated and GA-treated plants of the two mutants, *gai-t6 rga24* and *gai-t6 rga24 ga1-3* are located close to the non-GA-treated wt plants in the score plot.



A

Mass spectrum	Identified compound	*ga1-3*	*gai-t6 ga1-3*	*wt*
UPSC10028_GCTOF_Ath_Leaves_RI_1280	Phosphoric acid[a]	nn/n	n/n	pp/p
UPSC10064_GCTOF_Ath_Leaves_RI_1435	UNKNOWN CLASS	pp/pp*	pp/pp*	nn/pp*
UPSC10074_GCTOF_Ath_Leaves_RI_1476	EITTMS_148006-101-1_MST_1480.5_[NA][b]/ CARBOHYDRATE	pp/pp	pp/nn	nn/nn
UPSC10092_GCTOF_Ath_Leaves_RI_1590	CARBOHYDRATE/PENTOSE	nn/n *	n/nn*	p/nn*
UPSC10103_GCTOF_Ath_Leaves_RI_1624	UNKNOWN CLASS	nn/n *	n/n *	p/n *
UPSC10105_GCTOF_Ath_Leaves_RI_1635	UNKNOWN CLASS	nn/n	nn/nn	pp/ p
UPSC10110_GCTOF_Ath_Leaves_RI_1653	CARBOHYDRATE	n/nn*	n/nn*	p/nn*
UPSC10137_GCTOF_Ath_Leaves_RI_1752	Glyceraldehyde[a]	pp/p	p/n	nn/nn
UPSC10142_GCTOF_Ath_Leaves_RI_1781	EIQTMS_180013-101-1_MST_1804_[NA][b]/ UNKNOWN CLASS	pp/p	p/nn	nn/nn
UPSC10149_GCTOF_Ath_Leaves_RI_1826	POLYHYDROXY	pp/nn	p/nn	nn/nn
UPSC10150_GCTOF_Ath_Leaves_RI_1833	POLYHYDROXY	pp/n	p/nn	nn/nn
UPSC10158_GCTOF_Ath_Leaves_RI_1886	Glucose[a]	pp/pp*	p/pp*	nn/pp*
UPSC10159_GCTOF_Ath_Leaves_RI_1898	CARBOHYDRATE	pp/pp*	p/p *	nn/pp*
UPSC10199_GCTOF_Ath_Leaves_RI_2144	CARBOHYDRATE	nn/nn	nn/nn	nn/pp
UPSC10204_GCTOF_Ath_Leaves_RI_2180	2-O-Glycerol-beta-D-galactopyranoside[a]	pp/pp	nn/nn	nn/nn
UPSC10207_GCTOF_Ath_Leaves_RI_2200	EITTMS_221004-101-2_MST_2214.3_[NA][b]/ CARBOHYDRATE PHOSPHATE	n/n	n/n	n/p
UPSC10219_GCTOF_Ath_Leaves_RI_2292	Fructose-6-phosphate[a]	n/n	nn/n	n/n
UPSC10221_GCTOF_Ath_Leaves_RI_2301	Galactopyranoside, 1-O-methyl-, 2,3,4,6-tetrakis-O-(trimethylsilyl)-, alpha-d	pp/pp	nn/p	nn/n
UPSC10231_GCTOF_Ath_Leaves_RI_2355	UNKNOWN CLASS	n /nn	nn/n	n/nn
UPSC10254_GCTOF_Ath_Leaves_RI_2517	DISACCHARIDE	nn/n	pp/nn	pp/pp
UPSC10262_GCTOF_Ath_Leaves_RI_2604	GA_4[a]	pp/pp	pp/pp	pp/pp
UPSC10268_GCTOF_Ath_Leaves_RI_2778	MYO-INOSITOL-PHOSPHATE LIKE	nn/nn	n/nn	p/pp
UPSC10285_GCTOF_Ath_Leaves_RI_2966	Galactinol[a]	p/p	n/n	n/p
UPSC10295_GCTOF_Ath_Leaves_RI_3112	Beta-d-glucopyranose, 2,3,4,6-tetrakis-o-(trimethylsilyl)-, 1-(trimethylsilyl)-1h-indole-3-acetate	nn/nn	pp/nn	pp/pp

B

Mass spectrum	Identified compound	*ga1-3*	*gai-t6 ga1-3*	*wt*
UPSC10014_GCTOF_Ath_Leaves_RI_1224	UNKNOWN CLASS	nn/n	nn/n	
UPSC10043_GCTOF_Ath_Leaves_RI_1362	Serine[a]	n/n	n/n	
UPSC10059_GCTOF_Ath_Leaves_RI_1419	EITTMS_144004-101-2_MST_1436.4_[NA][b]/ UNKNOWN CLASS	n/nn	n/n	
UPSC10083_GCTOF_Ath_Leaves_RI_1549	UNKNOWN CLASS	nn/nn	n/n	
UPSC10096_GCTOF_Ath_Leaves_RI_1602	UNKNOWN CLASS	nn/n	nn/n	
UPSC10112_GCTOF_Ath_Leaves_RI_1658	UNKNOWN CLASS	pp/pp	nn/nn	
UPSC10118_GCTOF_Ath_Leaves_RI_1685	UNKNOWN CLASS	nn/nn	nn/n	
UPSC10126_GCTOF_Ath_Leaves_RI_1702	UNKNOWN CLASS	pp/nn	n/pp	
UPSC10153_GCTOF_Ath_Leaves_RI_1836	UNKNOWN CLASS	nn/nn	nn/nn	
UPSC10156_GCTOF_Ath_Leaves_RI_1870	UNKNOWN CLASS	pp/nn	n/nn	
UPSC10162_GCTOF_Ath_Leaves_RI_1926	GLUCONIC ACID LACTONE-LIKE	nn/p	nn/n	
UPSC10205_GCTOF_Ath_Leaves_RI_2185	CARBO-PHOSPHATE	nn/nn	nn/nn	
UPSC10214_GCTOF_Ath_Leaves_RI_2248	CARBOHYDRATE	nn/n	n/n	
UPSC10216_GCTOF_Ath_Leaves_RI_2264	EITTMS_228001-101-1_MST_2277_[NA][b]/ UNKNOWN CLASS	p/pp	n/n	
UPSC10224_GCTOF_Ath_Leaves_RI_2323	HEXOSE PHOSPHATE (ALDOSE)	nn/nn	nn/n	
UPSC10240_GCTOF_Ath_Leaves_RI_2398	Myo-inositol-1-phosphate[a]	n/n	n/n	
UPSC10260_GCTOF_Ath_Leaves_RI_2578	1-monohexadecanoylglycerol[a]	p/n	nn/n	
UPSC10274_GCTOF_Ath_Leaves_RI_2837	CARBOHYDRATE	n/nn	nn/n	

C

Mass spectrum	Identified compound	*ga1-3*	*gai-t6 ga1-3*	*wt*
UPSC10009_GCTOF_Ath_Leaves_RI_1209	UNKNOWN CLASS	n/n		p/n
UPSC10060_GCTOF_Ath_Leaves_RI_1422	CARBOHYDRATE	n/pp		n/n
UPSC10108_GCTOF_Ath_Leaves_RI_1648	Xylose[a]	pp/pp		n/n
UPSC10128_GCTOF_Ath_Leaves_RI_1710	CARBOHYDRATE/INOSE-LIKE	pp/pp		n/n
UPSC10155_GCTOF_Ath_Leaves_RI_1863	Fructose[a]	pp/pp*		n/p *
UPSC10160_GCTOF_Ath_Leaves_RI_1904	UNKNOWN CLASS	nn/nn*		pp/nn*
UPSC10165_GCTOF_Ath_Leaves_RI_1951	CARBOHYDRATE	pp/pp		p/p
UPSC10167_GCTOF_Ath_Leaves_RI_1968	EITTMS_199004-101-1_MST_1987.3_[NA][b]	pp/pp*		nn/p *
UPSC10182_GCTOF_Ath_Leaves_RI_2018	Beta-d-methylglucopyranoside[a]	pp/pp*		n/pp*
UPSC10191_GCTOF_Ath_Leaves_RI_2078	Inositol[a]	pp/pp*		n/p *
UPSC10259_GCTOF_Ath_Leaves_RI_2574	Nicotianamine[a]	nn/nn		n/nn
UPSC10263_GCTOF_Ath_Leaves_RI_2709	PHENYLIC COMPOUND + SUGAR	nn/nn		pp/pp
UPSC10264_GCTOF_Ath_Leaves_RI_2720	Maltose[a]	nn/nn		pp/pp
UPSC10273_GCTOF_Ath_Leaves_RI_2833	CARBOHYDRATE	n/nn		pp/p
UPSC10300_GCTOF_Ath_Leaves_RI_3220	Turanose[a]	nn/nn		p/pp

D

Mass spectrum	Identified compound	*ga1-3*	*gai-t6 ga1-3*	*wt*
UPSC10035_GCTOF_Ath_Leaves_RI_1319	Succinic acid[a]		p/n	n/n
UPSC10037_GCTOF_Ath_Leaves_RI_1332	Glyceric acid[a]		n/p	n/p
UPSC10050_GCTOF_Ath_Leaves_RI_1384	Threonine[a]		nn/nn	n/p
UPSC10056_GCTOF_Ath_Leaves_RI_1410	UNKNOWN CLASS		n/nn	n/n
UPSC10057_GCTOF_Ath_Leaves_RI_1411	UNKNOWN CLASS		n/n	p/n
UPSC10065_GCTOF_Ath_Leaves_RI_1436	UNKNOWN CLASS		n/n	n/nn
UPSC10075_GCTOF_Ath_Leaves_RI_1486	Malic acid[a]		nn/nn	n/n
UPSC10084_GCTOF_Ath_Leaves_RI_1555	Threonic acid[a]		pp/nn	n/n
UPSC10093_GCTOF_Ath_Leaves_RI_1593	UNKNOWN CLASS		pp/p	nn/nn
UPSC10100_GCTOF_Ath_Leaves_RI_1616	Glutamic acid[a]		nn/nn	nn/n
UPSC10125_GCTOF_Ath_Leaves_RI_1700	UNKNOWN CLASS		n/n	n/nn
UPSC10129_GCTOF_Ath_Leaves_RI_1711	PENTOL (LIKE XYLITOL/ARABITOL)		nn/n	nn/nn
UPSC10131_GCTOF_Ath_Leaves_RI_1716	Ribitol[a]		p/pp	n/n
UPSC10145_GCTOF_Ath_Leaves_RI_1803	Shikimic acid[a]		n/nn	n/n
UPSC10147_GCTOF_Ath_Leaves_RI_1810	Citric acid[a]		nn/nn	nn/nn
UPSC10163_GCTOF_Ath_Leaves_RI_1937	Ascorbic acid[a]		nn/n	pp/p
UPSC10176_GCTOF_Ath_Leaves_RI_1995	UNKNOWN CLASS		nn/nn	pp/nn
UPSC10192_GCTOF_Ath_Leaves_RI_2095	EITTMS_211001-101-2_MST_2105.7_[B12][b]/ HEXOSE		n/n	n/n
UPSC10252_GCTOF_Ath_Leaves_RI_2488	EITTMS_251001-101-1_MST_2505_[NA][b]/ UNKNOWN CLASS CARBOHYDRATE		n/n	n/n
UPSC10312_GCTOF_Ath_Leaves_RI_2623	Sucrose[a]		n/n	nn/n
UPSC10267_GCTOF_Ath_Leaves_RI_2763	EIQTMS_278005-101-1_MST_2783.3_[NA][b]/ UNKNOWN CLASS		nn/n	p/p
UPSC10269_GCTOF_Ath_Leaves_RI_2790	UNKNOWN CLASS		nn/n	nn/pp
UPSC10296_GCTOF_Ath_Leaves_RI_3113	CARBOHYDRATE		nn/nn	nn/p
UPSC10308_GCTOF_Ath_Leaves_RI_3369	Raffinose[a]		p/nn	nn/pp

Figure 14. Venn diagram and identity of metabolites whose levels changed significantly (p=5%) following GA_4 treatment in *ga1-3, gai-t6 rga-24 ga1-3* and WT. Peaks are labelled according to the names of their equivalents in the UPSC mass spectra database (UPSC##_GCTOF_Species_RI#). "p" and "n" denote increases and decreases, respectively, in the metabolic response following GA treatment. The first and second letters before the backslash indicate the difference between the control samples and the GA-treated samples at 24 hours and 48 hours, respectively. Effects larger and smaller than 37% are indicated by "pp and nn" respectively. [a]The metabolites were identified as methoxyamine-trimethylsilyl derivatives using an in-house mass spectra library or the mass spectra library hosted by the Max Planck Institute in Golm (http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html, 19-August-2005). [b]Identity to a mass spectrum in the mass spectra library of the Max Planck Institute in Golm. Asterisks (*) indicate examples of different pulse effects (time responses) between the three genotypes.

The addition of GA_4 to the *ga1-3* plants resulted in a change in the metabolite composition, causing the samples to shift close to not only the *gai-t6 rga24* and *gai-t6 rga24 ga1-3* mutants, but also to the non-treated wt plants in the score plot. Similarly, the *gai-t6 ga1-3* plants moved towards the double mutants, triple mutants and the non-treated wt plants after GA_4 treatment. The *ga1-3* mutant moved further

in the score plot that any other genotype following GA treatment. In the score plot the *gai-t6 rga24* and *gai-t6 rga24 ga1-3* mutants show weak responses to GA_4 treatment in comparison with the other genotypes. After treatment the wt plants moved in a different direction to *ga1-3* and *gai-t6 ga1-3*. The *gai t6, rga24, sly1-10* and *sly1-10* mutants moved towards each other in opposite directions after treatment. The *gai* mutant moved toward the GA-treated wt plants.

In earlier work, application of 10^{-5} M GA_4 has restored the severe dwarf *ga1-3* to wt phenotype within seven days (King, Moritz & Harberd, 2001). We can see from the score plot that the metabolome of the *ga1-3* plant becomes similar to wt after GA_4-treatment within 48 hours. This is not surprising as *ga1-3* is more or less lacking in GAs, so treatment with GA_4 should rapidly restore its phenotype. The two DELLA mutants *gai-t6 rga24* and *gai-t6 rga24 ga1-3* have similar phenotypes to wt since both RGA and GAI are the major repressors during vegetative growth (Dill & Sun, 2001; King, Moritz & Harberd, 2001; Tyler *et al.*, 2004). It has also been shown that application of GA_4 increases the hypocotyl length not only for wt but also, albeit to a smaller degree, for the DELLA mutants *gai-t6 ga1-3, rga24 ga1-3* and *gai-t6 rga24 ga1-3* (King, Moritz & Harberd, 2001). On a metabolic level this effect does not seem to be so dramatic, but a similarity between them and the wt control plants can be seen. On the other hand, the shift induced by GA_4 in the WT is the second largest in the score plot. The change of position for the *gai* mutant may be due to the putative effect of GA_4 on degradation of the RGA protein. The changed positions of the *gai t6, rga24, sly1-10* and *sly1-10* mutants are more difficult to explain according to the literature. Of the metabolites significantly affected by the GA_4 treatment or the synergistic effects of GA_4 treatment and time, around half could be identified using libraries and retention index information (Schauer *et al.*, 2005). To further investigate the effect of adding GA on *gai-t6 ga1-3, rga24 ga1-3* and wt genotypes, the metabolites that significantly changed were investigated. The compounds, their unique UPSC names and retention indices (based on the C_{12}-C_{40} series of n-alkanes) are listed in Figure 14. The annotations p and n stand for increases and decreases in metabolic responses after GA treatment. The first letter before the backslash corresponds to the effect between the control samples and the GA-treated samples at 24 hours, and the second at 48 hours. Effects larger and smaller than 37% are represented by pp and nn, respectively. 37% is the average effect for the GA treatment for the 284 peaks. The identified metabolites are mainly sugars and sugar phosphates, indicating that GA_4 treatment had a general effect on primary metabolism.

The GA treatment also affected the IAA levels in the mutants (Figure 15). For the control samples the levels were similar, apart from being high in *gai-t6 ga1-3*, which is surprising. After applying active GA the levels of IAA declined in the *gai-t6 ga1-3* and *ga1-3* mutants. This effect was significant ($p<1\%$) according to OLS calibration using MODDE. Similarly to the results of the metabolite profiling studies (Figure 13), the *gai-t6 ga1-3* and *ga1-3* mutants became more similar to WT, and the levels of IAA declined to levels similar to those of the non-GA-treated WT plants (Figure 15). The effects of GA application were rapid in both of these mutants, as the levels for IAA were almost constant after 24 h of treatment. Comparing with changes in the metabolite levels that significantly changed after The levels of a number of metabolites significantly changed following GA treatment in the *ga1-3*, *gai-t6 ga1-3* and WT plants, and showed similar trends to the changes in IAA levels, especially

carbohydrates (Figure 15).This trend relates to different pulse effects after adding the GA_4 to the three mutants. *ga1-3* showed a faster response to the GA treatment than *gai-t6 ga1-3* and wt. For example, the levels of glucose increased markedly within 24 of treatment in *ga1-3* (>37%) and continued to be high after 48 hours. A slower effect can be seen for *gai-t6 ga1-3*, where the levels had increased (<37%) after 24 h and continued to increase thereafter (>37% at 48 h). For wt the glucose levels were lower than average for the GA-treated plant at 24 h (<-37%) but were greater at 48 h (>37%). Similar trends in the opposite direction can also be seen for some metabolites, i.e. their levels decreased rather than increased more rapidly in *ga1-3* than in *gai-t6 ga1-3* and wt. In recent years it has been shown that levels of GA and IAA can be restored by IAA application (Ross *et al.*, 2000; Wolbang *et al.*, 2004). Our results indicate that there is a balance between the hormone groups and after addition of GA to GA-deficient mutants the levels of IAA are restored to WT levels. Auxin has been proposed to be a positive regulator of *AtGA3ox1* (Figure 8), which catalyses the conversion of GA_9 to active GA_4 (Wolbang *et al.*, 2004). The fast reduction of IAA and the delay in the effect on the metabolome of GA_4 treatment suggest that IAA and GA levels are tightly controlled by each other's levels.

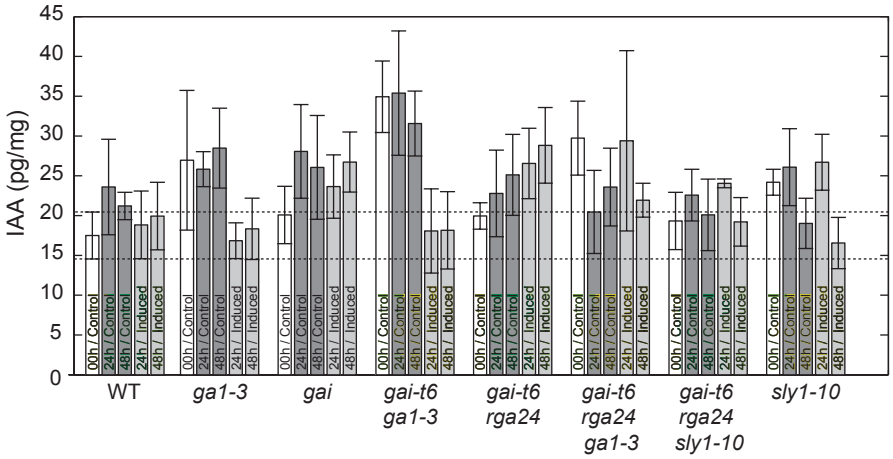


Figure 15. IAA levels in GA mutants following treatment by GA_4 after 24 and 48 hours. Additional control measurements were performed at 0, 24 and 48 hours. The error bars represent standard deviations.

We have shown that a metabolic analysis can be performed in a comprehensive way. The analysis of a large series of plant samples and data processing was completed within a week with minimal labour intensive work. The processing method presented in Paper III provides an efficient data processing tool, generating results that are biologically interpretable. The variation of interest, the time effect after GA treatment, could be separated from the other introduced variation by comparing the non-normalized weight vectors between different PLS models.

Conclusions and Future Plans

In the work underlying this thesis I have developed methods that can be used to compare metabolic profiles of plant samples. I have shown that reliable protocols for metabolomic analysis can be developed using few experiments according to DOE. The method presented here is rapid, and involves steps that can be quite easily automated. The use of GC/TOFMS systems allows fast, high through-put analysis (90 samples per 24 h). The use of hierarchical processing methods for GC/MS to resolve complex mixtures also allows fast comparisons of complex samples. The data processing methods developed here provide useful tools for generating biologically interpretable results. The methods were used to analyse metabolites in a number *Arabidopsis* mutants lacking GAs and/or parts of the GA-signalling pathways after GA application. The analysis of a large series of plant samples and subsequent data processing were completed within a week, with minimal labour intensive work.

To exploit the potential of metabolomics in plant sciences as fully as possible there is a need not only to increase the number of metabolites detected and identified, but also to improve our ability to interpret the results. This implies that there is need for new instrumental and multivariate techniques. In the "omics" fields there is a need to assess many different types of biological treatments and replicates, so samples have to be analyzed in a high through-put manner. UPLC™ and CE (Soga *et al.*, 2003; Sato *et al.*, 2004) have shown great separation efficiency (Shen *et al.*, 2005; Wilson *et al.*, 2005a) and proved, in combination with mass spectrometry, to provide good alternatives for detecting new metabolites. In addition, ion mobility spectrometry (IMS; Tang *et al.*, 2005) in combination with mass spectrometry would be a suitably fast method for separating mixtures. For improving the handling of MS-data, further developments of AR compression and Multivariate Curve Resolution for GC are required, as are improved comparisons for LC/MS, in similar ways to those described by Jonsson *et al.* (2005). The robustness of the methods must be investigated, but the scope for processing external samples should also be examined. Better GC/MS data pre-treatment methodologies must also be developed, such as background subtraction and alignment, that have not been addressed in this thesis.

Current analytical protocols, even those including a number of extractions, purification and separation steps, can be more easily automated than previous methodologies. Nevertheless, in order to analyze as many metabolites as possible, a number of analytical issues have to be addressed. In order to develop analytical protocols for profiling techniques new strategies for Design of experiment (DOE) have to be considered. The responses are complex, so compromises have to be made in the final protocol. However, if the process is automated the number of experiments is not limited, or at least it is less limited than it generally was in previous applications of DOE.

Comparison of mass spectra and retention times against a mass spectra database containing library compounds is important for identification, but will not provide coverage of all metabolites since it is impossible to generate relevant information on all of the metabolites synthesized by plants. Alternative approaches are needed. For instance, the molecular information can be transformed into discrete data by describing compounds with chemical descriptors. This information, together with mass spectra information, can be related to experimental data. Delimited calibration

models, using for example PLS, can then be compiled for known metabolites and used to predict or classify unknown compounds. We propose that Quantitative Structure-Retention Relationships (QSRR; Nord, Fransson & Jacobsson, 1998) could be used as identification tools for metabolomics. The identity of metabolites could then be confirmed, by comparing the empirical information with retention prediction models. This can be useful when the candidate metabolites are either difficult to obtain in large enough quantities for NMR identification, or are too time consuming to synthesize.

Given the differences in delay and metabolomic effects of GA_4 treatment on the GA mutants, and the fast reduction in IAA levels in the *ga1-3* and *gai-t6 ga1-3* mutants, it would be interesting to study the levels of IAA after shorter time periods than 24 hours, as would studying the reductions in IAA levels in the *ga1-3* and *gai-t6 ga1-3* mutants to WT levels over longer periods of rime. This would determine whether or not the IAA levels continue to match WT levels after application of GA_4. Another task that must be undertaken to further elucidate the crosstalk between the hormones is to identify unknown metabolites.

References

Adams, MA., Chen, ZL., Landman, P. & Colmer, TD. 1999. Simultaneous determination by capillary gas chromatography of organic acids, sugars, and sugar alcohols in plant tissue extracts as their trimethylsilyl derivatives. *Analytical Biochemistry 266*, 77-84.

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 Biotechnology 21*, 692-696.

Alvey, L. & Harberd, NP. 2005. DELLA proteins: integrators of multiple plant growth regulatory inputs? *Physiologia Plantarum 123*, 153-160.

Andersson, A., Keskitalo, J., Sjödin, A., Bhalerao, R., Sterky, F., Wissel, K., Tandre, K., Aspeborg, H., Moyle, R., Ohmiya, Y., Bhalerao, R., Brunner, A., Gustafsson, P., Karlsson, J., Lundeberg, J., Nilsson, O., Sandberg, G., Strauss, S., Sundberg, B., Uhlen, M., Jansson, S. & Nilsson, P. 2004. A transcriptional timetable of autumn senescence. *Genome Biology 5*,

Antti, H., Ebbels, TMD., Keun, HC., Bollard, ME., Beckonert, O., Lindon, JC., Nicholson, JK. & Holmes, E. 2004. Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects. *Chemometrics And Intelligent Laboratory Systems 73*, 139-149.

Asres, DD. & Perreault, H. 1997. Monosaccharide permethylation products for gas chromatography mass spectrometry: how reaction conditions can influence isomeric ratios. *Canadian Journal of Chemistry-Revue Canadienne De Chimie 75*, 1385-1392.

Barclay, VJ., Bonner, RF. & Hamilton, IP. 1997. Application of wavelet transforms to experimental spectra: Smoothing, denoising, and data set compression. *Analytical Chemistry 69*, 78-90.

Benthin, B., Danz, H. & Hamburger, M. 1999. Pressurized liquid extraction of medicinal plants. *Journal Of Chromatography A 837*, 211-219.

Bezemer, E. & Rutan, S. 2001. Study of the hydrolysis of a sulfonylurea herbicide using liquid chromatography with diode array detection and mass spectrometry by three-way multivariate curve resolution-alternating least squares. *Analytical Chemistry 73*, 4403-4409.

Bino, RJ., Hall, RD., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, BJ., Mendes, P., Roessner-Tunali, U., Beale, MH., Trethewey, RN., Lange, BM., Wurtele, ES. & Sumner, LW. 2004. Potential of metabolomics as a functional genomics tool. *Trends In Plant Science 9*, 418-425.

Blau, K. & Halket, JM. 1993. *Handbook of Derivatives for Chromatography*. editon. John Wiley & Sons Ltd. Chichester. pp

Bolle, C. 2004. The role of GRAS proteins in plant signal transduction and development. *Planta 218*, 683-692.

Bro, R. 2003. Multivariate calibration - What is in chemometrics for the analytical chemist? *Analytica Chimica Acta 500*, 185-194.

Broeckling, CD., Huhman, DV., Farag, MA., Smith, JT., May, GD., Mendes, P., Dixon, RA. & Sumner, LW. 2005. Metabolic profiling of Medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *Journal Of Experimental Botany 56*, 323-336.

Carlson, R. & Carlson, JE. 2005. *Design and Optimization in Organic Synthesis: Second Revised and Enlarged Edition (Data Handling in Science and Technology)*. editon. Elsevier Science. Amsterdam. 596 pp

Cech, NB. & Enke, CG. 2001. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews 20*, 362-387.

Choi, HK., Choi, YH., Verberne, M., Lefeber, AWM., Erkelens, C. & Verpoorte, R. 2004. Metabolic fingerprinting of wild type and transgenic tobacco plants by H-1 NMR and multivariate analysis technique. *Phytochemistry 65*, 857-864.

Cook, D., Fowler, S., Fiehn, O. & Thomashow, MF. 2004. A prominent role for the CBF cold response pathway in configuring the low-temperature metabolome of Arabidopsis. *Proceedings Of The National Academy Of Sciences Of The United States Of America 101*, 15243-15248.

Curtius, H-C., Muller, M. & Völlmin, JA. 1968. Studies of the Ring Structures of Ketoses by Means of Gaschromatography and Mass Specstroscopy. *J. Chromatography 37*, 216-224.

Davies, PJ. 2004. *Plant Hormones: physiology, biochemistry and molecular biology*. 3rd editon. Kluwer Academic Publishers. Dordrecht, The Netherlands. 750 pp

de Aguiar, PF., Bourguignon, B., Khots, MS., Massart, DL. & PhanThanLuu, R. 1995. D-optimal designs. *Chemometrics And Intelligent Laboratory Systems 30*, 199-210.

de Hoffmann, E. & Stroobant, V. 2002. *Mass Spectrometry: Principles and Applications*. 2nd editionrd editon. John Wiley & Sons. New York. 420 pp

Desbrosses, GG., Kopka, J. & Udvardi, MK. 2005. Lotus japonicus metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. *Plant Physiology 137*, 1302-1318.

Dill, A. & Sun, TP. 2001. Synergistic derepression of gibberellin signaling by removing RGA and GAI function in Arabidopsis thaliana. *Genetics 159*, 777-785.

Dumouchel, W. & Jones, B. 1994. A Simple Bayesian Modification of D-Optimal Designs to Reduce Dependence on an Assumed Model. *Technometrics 36*, 37-47.

Dunn, WB. & Ellis, DI. 2005. Metabolomics: Current analytical platforms and methodologies. *Trac-Trends In Analytical Chemistry 24*, 285-294.

Duran, AL., Yang, J., Wang, LJ. & Sumner, LW. 2003. Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics 19*, 2283-2293.

Edlund, A., Eklöf, S., Sundberg, B., Moritz, T. & Sandberg, G. 1995. A Microscale Technique for Gas-Chromatography Mass-Spectrometry Measurements of Picogram Amounts of Indole-3-Acetic-Acid in Plant-Tissues. *Plant Physiology 108*, 1043-1047.

Edwards, JL. & Kennedy, RT. 2005. Metabolomic analysis of eukaryotic tissue and prokaryotes using negative mode MALDI time-of-flight mass spectrometry. *Analytical Chemistry 77*, 2201-2209.

Efron, B. 1986. Jackknife, Bootstrap and Other Resampling Methods in Regression-Analysis - Discussion. *Annals of Statistics 14*, 1301-1304.

Eide, I., Neverdal, G., Thorvaldsen, B., Shen, HL., Grung, B. & Kvalheim, O. 2001. Resolution of GC-MS data of complex PAC mixtures and regression modeling of mutagenicity by PLS. *Environmental Science & Technology 35*, 2314-2318.

Evershed, RP. 1993. *Advances in Silylation*. Ed John Wiley & Sons Ltd. Chichester. pp 58-59

Fenn, JB., Mann, M., Meng, CK., Wong, SF. & Whitehouse, CM. 1989. Electrospray Ionization For Mass-Spectrometry Of Large Biomolecules. *Science 246*, 64-71.

Fernie, AR., Geigenberger, P. & Stitt, M. 2005. Flux an important, but neglected, component of functional glenomics. *Current Opinion In Plant Biology 8*, 174-182.

Fiehn, O. 2002. Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology 48*, 155-171.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, RN. & Willmitzer, L. 2000a. Metabolite profiling for plant functional genomics. *Nature Biotechnology 18*, 1157-1161.

Fiehn, O., Kopka, J., Trethewey, RN. & Willmitzer, L. 2000b. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical Chemistry 72*, 3573-3580.

Fleet, CM. & Sun, TP. 2005. A DELLAcate balance: the role of gibberellin in plant

morphogenesis. *Current Opinion In Plant Biology 8*, 77-85.
Fraga, CG., Prazen, BJ. & Synovec, RE. 2001. Objective data alignment and chemometric analysis of comprehensive two-dimensional separations with run-to-run peak shifting on both dimensions. *Analytical Chemistry 73*, 5833-5840.
Fu, XD. & Harberd, NP. 2003. Auxin promotes Arabidopsis root growth by modulating gibberellin response. *Nature 421*, 740-743.
Gamiz-Gracia, L. & de Castro, MDL. 2000. Continuous subcritical water extraction of medicinal plant essential oil: comparison with conventional techniques. *Talanta 51*, 1179-1185.
Garratt, LC., Linforth, R., Taylor, AJ., Lowe, KC., Power, JB. & Davey, MR. 2005. Metabolite fingerprinting in transgenic lettuce. *Plant Biotechnology Journal 3*, 165-174.
Gemperline, P. 1984. A priori estimates of the elution profiles of the pure components in overlapped liquid chromatography peaks using target factor analysis. *J Chem Info Computer Sci 24*, 206-212.
German, JB., Roberts, MA. & Watkins, SM. 2003. Personal metabolomics as a next generation nutritional assessment. *Journal Of Nutrition 133*, 4260-4266.
Gomi, K. & Matsuoka, M. 2003. Gibberellin signalling pathway. *Current Opinion In Plant Biology 6*, 489-493.
Goodacre, R. 2005. Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *Journal Of Experimental Botany 56*, 245-254.
Grande, BV. & Manne, R. 2000. Use of convexity for finding pure variables in two-way data from mixtures. *Chemometrics And Intelligent Laboratory Systems 50*, 19-33.
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 Communications In Mass Spectrometry 13*, 279-284.
Harrigan, GG. & Goodacre, R. 2003. *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. editon. Kluwer Academic Publishers. London, UK. 352 pp
Hedden, P. & Phillips, AL. 2000. Gibberellin metabolism: new insights revealed by the genes. *Trends In Plant Science 5*, 523-530.
Herbert, C, G. & Johnstone, RAW. 2002. *Mass Spectrometry Basics*. editon. CRC Press. New York. 496 pp
Horning, EC. & Horning, MG. 1971. Human metabolic profiles obtained by GC and GC/MS. *Journal of Chromatographic Science 9*, 129–140.
Huhman, DV. & Sumner, LW. 2002. Metabolic profiling of saponins in Medicago sativa and Medicago truncatula using HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry 59*, 347-360.
Huie, CW. 2002. A review of modern sample-preparation techniques for the extraction and analysis of medicinal plants. *Analytical And Bioanalytical Chemistry 373*, 23-30.
Höskuldsson, A. 1995. A Combined Theory For PCA and PLS. *Journal Of Chemometrics 9*, 91-123.
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. *Analytical Chemistry 75*, 4784-4792.
Jackson, EJ. 1991. *A User's Guide to Principal Components*. editon. Wiley-Interscience. New York. 569 pp
Jiang, TH., Liang, Y. & Ozaki, Y. 2004. Principles and methodologies in self-modeling curve resolution. *Chemometrics And Intelligent Laboratory Systems 71*, 1-12.
Johansen, HN., Glitso, V. & Knudsen, KEB. 1996. Influence of extraction solvent and temperature on the quantitative determination of oligosaccharides from plant materials by high-performance liquid chromatography. *Journal Of Agricultural And Food Chemistry 44*, 1470-1474.

Johnson, ME. & Nachtsheim, CJ. 1983. Some Guidelines for Constructing Exact D-Optimal Designs on Convex Design Spaces. *Technometrics 25*, 271-277.
Jonsson, P., Bruce, SJ., Moritz, T., Trygg, J., Sjöström, M., Plumb, R., Granger, J., Maibaum, E., Nicholson, JK., Holmes, E. & Antti, H. 2005. Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. *Analyst 130*, 701-707.
Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Research 30*, 42-46.
Karjalainen, EJ. 1989. The Spectrum Reconstruction Problem - Use of Alternating Regression for Unexpected Spectral Components in 2-Dimensional Spectroscopies. *Chemometrics And Intelligent Laboratory Systems 7*, 31-38.
Kaufmann, B. & Christen, P. 2002. Recent extraction techniques for natural products: Microwave- assisted extraction and pressurised solvent extraction. *Phytochemical Analysis 13*, 105-113.
Kebarle, P. & Peschke, M. 2000. On the mechanisms by which the charged droplets produced by electrospray lead to gas phase ions. *Analytica Chimica Acta 406*, 11-35.
King, KE., Moritz, T. & Harberd, NP. 2001. Gibberellins are not required for normal stem growth in Arabidopsis thaliana in the absence of GAI and RGA. *Genetics 159*, 767-776.
King, R., Bonfiglio, R., Fernandez-Metzler, C., Miller-Stein, C. & Olah, T. 2000. Mechanistic investigation of ionization suppression in electrospray ionization. *Journal Of The American Society For Mass Spectrometry 11*, 942-950.
Kvalheim, OM. 1992. The Latent Variable. *Chemometrics And Intelligent Laboratory Systems 14*, 1-3.
Kvalheim, OM. & Liang, YZ. 1992. Heuristic Evolving Latent Projections - Resolving 2-Way Multicomponent Data.1. Selectivity, Latent-Projective Graph, Datascope, Local Rank, and Unique Resolution. *Analytical Chemistry 64*, 936-946.
Lahner, B., Gong, JM., Mahmoudian, M., Smith, EL., Abid, KB., Rogers, EE., Guerinot, ML., Harper, JF., Ward, JM., McIntyre, L., Schroeder, JI. & Salt, DE. 2003. Genomic scale profiling of nutrient and trace elements in Arabidopsis thaliana. *Nature Biotechnology 21*, 1215-1221.
Lange, BM. & Ghassemian, M. 2005. Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. *Phytochemistry 66*, 413-451.
Lawrence, E. 1995. *Henderson's dictionary of biological terms*. 11rd editon. Longman Scientific & Technical. Essex, England. 693 pp
Le Gall, G., Colquhoun, IJ., Davis, AL., Collins, GJ. & Verhoeyen, ME. 2003a. Metabolite profiling of tomato (Lycopersicon esculentum) using H-1 NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. *Journal Of Agricultural And Food Chemistry 51*, 2447-2456.
Le Gall, G., DuPont, MS., Mellon, FA., Davis, AL., Collins, GJ., Verhoeyen, ME. & Colquhoun, IJ. 2003b. Characterization and content of flavonoid glycosides in genetically modified tomato (Lycopersicon esculentum) fruits. *Journal Of Agricultural And Food Chemistry 51*, 2438-2446.
Leavens, WJ., Lane, SJ., Carr, RM., Lockie, AM. & Waterhouse, I. 2002. Derivatization for liquid chromatography/electrospray mass spectrometry: synthesis of tris(trimethoxyphenyl)phosphonium compounds and their derivatives of amine and carboxylic acids. *Rapid Communications In Mass Spectrometry 16*, 433-441.
Lee, SC., Cheng, H., King, KE., Wang, WF., He, YW., Hussain, A., Lo, J., Harberd, NP. & Peng, JR. 2002. Gibberellin regulates Arabidopsis seed germination via RGL2, a GAI/RGA-like gene whose expression is up-regulated following imbibition. *Genes & Development 16*, 646-658.
Lenz, EM., Bright, J., Knight, R., Wilson, ID. & Major, H. 2004. Cyclosporin A-induced changes in endogenous meta-bolites in rat urine: a metabonomic investigation using high field H-1 NMR spectroscopy, HPLC-TOF/MS and

chemometrics. *Journal Of Pharmaceutical And Biomedical Analysis 35*, 599-608.
Liang, YZ. & Kvalheim, OM. 1994. Diagnosis and Resolution of Multiwavelength Chromatograms by Rank Map, Orthogonal Projections and Sequential Rank Analysis. *Analytica Chimica Acta 292*, 5-15.
Liang, YZ. & Kvalheim, OM. 2001. Resolution of two-way data: theoretical background and practical problem-solving - Part 1: Theoretical background and methodology. *Fresenius Journal of Analytical Chemistry 370*, 694-704.
Little, JL. 1999. Artifacts in trimethylsilyl derivatization reactions and ways to avoid them. *Journal Of Chromatography A 844*, 1-22.
Ljung, K., Hull, AK., Celenza, J., Yamada, M., Estelle, M., Nonmanly, J. & Sandberg, G. 2005. Sites and regulation of auxin biosynthesis in Arabidopsis roots. *Plant Cell 17*, 1090-1104.
Lui, LH., Vikram, A., Abu-Nada, Y., Kushalappa, AC., Raghavan, GSV. & Al-Mughrabi, K. 2005. Volatile metabolic profiling for discrimination of potato tubers inoculated with dry and soft rot pathogens. *American Journal Of Potato Research 82*, 1-8.
Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, A., Pettersen, J. & Bergman, R. 1998. Experimental design and optimization. *Chemometrics And Intelligent Laboratory Systems 42*, 3-40.
Malinowski, ER. 1996. Automatic window factor analysis - A more efficient method for determining concentration profiles from evolutionary spectra. *Journal Of Chemometrics 10*, 273-279.
Malmquist, G. & Danielsson, R. 1994. Alignment of Chromatographic Profiles for Principal Component Analysis - a Prerequisite for Fingerprinting Methods. *Journal Of Chromatography A 687*, 71-88.
Mamyrin, BA. 2001. Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal Of Mass Spectrometry 206*, 251-266.
Manne, R. & Grande, BV. 2000. Resolution of two-way data from hyphenated chromatography by means of elementary matrix transformations. *Chemometrics And Intelligent Laboratory Systems 50*, 35-46.
Martens, H. & Martens, M. 2000. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Quality and Preference 11*, 5-16.
Martens, H. & Naes, T. 1992. *Multivariate Calibration*. 1rd editon. John Wiley & Sons. New York, USA. 438 pp
McGinnis, KM., Thomas, SG., Soule, JD., Strader, LC., Zale, JM., Sun, TP. & Steber, CM. 2003. The Arabidopsis SLEEPY1 gene encodes a putative F-box subunit of an SCF E3 ubiquitin ligase. *Plant Cell 15*, 1120-1130.
Meinke, DW., Cherry, JM., Dean, C., Rounsley, SD. & Koornneef, M. 1998. Arabidopsis thaliana: A model plant for genome analysis. *Science 282*, 662-682.
Mueller, LA., Zhang, PF. & Rhee, SY. 2003. AraCyc: A biochemical pathway database for Arabidopsis. *Plant Physiology 132*, 453-460.
Murashige, T. & Skoog, F. 1962. A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiologia Plantarum* 473-497.
Newton, RP., Brenton, AG., Smith, CJ. & Dudley, E. 2004. Plant proteome analysis by mass spectrometry: principles, problems, pitfalls and recent developments. *Phytochemistry 65*, 1449-1485.
Nicholson, JK., Connelly, J., Lindon, JC. & Holmes, E. 2002. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery 1*, 153-161.
Nord, LI., Fransson, D. & Jacobsson, SP. 1998. Prediction of liquid chromatographic retention times of steroids by three-dimensional structure descriptors and partial least squares modeling. *Chemometrics And Intelligent Laboratory Systems 44*, 257-269.
Nordström, A., Tarkowski, P., Tarkowska, D., Dolezal, K., Åstot, C., Sandberg, G. & Moritz, T. 2004. Derivatization for LC electrospray ionization-MS: A tool for

improving reversed-phase separation and ESI responses of bases, ribosides, and intact nucleotides. *Analytical Chemistry 76*, 2869-2877.
O'Hagan, S., Dunn, WB., Brown, M., Knowles, JD. & Kell, DB. 2005. Closed-loop, multiobjective optimization of analytical instrumentation: Gas chromatography/ time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Analytical Chemistry 77*, 290-303.
Okamoto, M., Takahashi, KI. & Doi, T. 1995. Sensitive Detection And Structural Characterization Of Trimethyl(P-Aminophenyl)Ammonium-Derivatized Oligosaccharides By Electrospray-Ionization Mass-Spectrometry And Tandem Mass-Spectrometry. *Rapid Communications In Mass Spectrometry 9*, 641-643.
Olszewski, N., Sun, TP. & Gubler, F. 2002. Gibberellin signaling: Biosynthesis, catabolism, and response pathways. *Plant Cell 14*, S61-S80.
Ong, ES. 2002. Chemical assay of glycyrrhizin in medicinal plants by pressurized liquid extraction (PLE) with capillary zone electrophoresis (CZE). *Journal of Separation Science 25*, 825-831.
Orth, HCJ., Rentel, C. & Schmidt, PC. 1999. Isolation, purity analysis and stability of hyperforin as a standard material from Hypericum perforatum L. *Journal of Pharmacy and Pharmacology 51*, 193-200.
Ozel, M., Gogus, F., Hamilton, JF. & Lewis, AC. 2005. Analysis of volatile components from Ziziphora taurica subsp taurica by steam distillation, superheated-water extraction, and direct thermal desorption with GCxGC-TOFMS. *Analytical And Bioanalytical Chemistry 382*, 115-119.
Pere-Trepat, E., Hildebrandt, A., Barcelo, D., Lacorte, S. & Tauler, R. 2004. Fast chromatography of complex biocide mixtures using diode array detection and multivariate curve resolution. *Chemometrics And Intelligent Laboratory Systems 74*, 293-303.
Pierce, AE. 1968. *Silylation of Organic Compounds*. editon. Pierce Chemical Company. Rockford. 487 pp
Riter, LS., Vitek, O., Gooding, KM., Hodge, BD. & Julian, RK. 2005. Statistical design of experiments as a tool in mass spectrometry. *JOURNAL OF MASS SPECTROMETRY 40*, 565-579.
Robertson, DG. 2005. Metabonomics in toxicology: A review. *Toxicological Sciences 85*, 809-822.
Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. & Fernie, AR. 2001. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell 13*, 11-29.
Roessner, U., Wagner, C., Kopka, J., Trethewey, RN. & Willmitzer, L. 2000. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal 23*, 131-142.
Ross, JJ., O'Neill, DP., Smith, JJ., Kerckhoffs, LHJ. & Elliott, RC. 2000. Evidence that auxin promotes gibberellin A(1) biosynthesis in pea. *Plant Journal 21*, 547-552.
Sanchez, FC., vandenBogaert, V., Rutan, SC. & Massart, DL. 1996. Multivariate peak purity approaches. *Chemometrics And Intelligent Laboratory Systems 34*, 139-171.
Sato, S., Soga, T., Nishioka, T. & Tomita, M. 2004. Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *Plant Journal 40*, 151-163.
Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., Lundgren, K., Roessner-Tunali, U., Forbes, MG., Willmitzer, L., Fernie, AR. & Kopka, J. 2005. GC-MS libraries for the rapid identification of metabolites in complex biological samples. *Febs Letters 579*, 1332-1337.
Schlichtherle-Cerny, H., Affolter, M. & Cerny, C. 2003. Hydrophilic interaction liquid chromatography coupled to electrospray mass spectrometry of small polar compounds in food analysis. *Analytical Chemistry 75*, 2349-2354.
Shen, YF., Zhang, R., Moore, RJ., Kim, J., Metz, TO., Hixson, KK., Zhao, R., Livesay, EA., Udseth, HR. & Smith, RD. 2005. Automated 20 kpsi RPLC-MS

and MS/MS with chromatographic peak capacities of 1000-1500 and capabilities in proteomics and metabolomics. *Analytical Chemistry 77*, 3090-3100.
Shurubor, YI., Paolucci, U., Krasnikov, BF., Matson, WR. & Kristal, BS. 2005. Analytical precision, biological variation, and mathematical normalization in high data density metabolomics. *Metabolomics 1*, 75.
Silverstone, AL., Mak, PYA., Martinez, EC. & Sun, TP. 1997. The new RGA locus encodes a negative regulator of gibberellin response in Arabidopsis thaliana. *Genetics 146*, 1087-1099.
Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M. & Nishioka, T. 2003. Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *Journal Of Proteome Research 2*, 488-494.
Steuer, R., Kurths, J., Fiehn, O. & Weckwerth, W. 2003. Observing and interpreting correlations in metabolomic networks. *Bioinformatics 19*, 1019-1026.
Streeter, JG. & Strimbu, CE. 1998. Simultaneous extraction and derivatization of carbohydrates from green plant tissues for analysis by gas-liquid chromatography. *Analytical Biochemistry 259*, 253-257.
Sumner, LW., Mendes, P. & Dixon, RA. 2003. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry 62*, 817-836.
Sun, TP. & Gubler, F. 2004. Molecular mechanism of gibberellin signaling in plants. *Annual Review Of Plant Biology 55*, 197-223.
Swain, SM. & Singh, DP. 2005. Tall tales from sly dwarves: novel functions of gibberellins in plant development. *Trends In Plant Science 10*, 123-129.
Tang, K., Shvartsburg, AA., Lee, HN., Prior, DC., Buschbach, MA., Li, FM., Tolmachev, AV., Anderson, GA. & Smith, RD. 2005. High-sensitivity ion mobility spectrometry/mass spectrometry using electrodynamic ion funnel interfaces. *Analytical Chemistry 77*, 3330-3339.
Tauler, R. 1995. Multivariate curve resolution applied to second order data. *Chemometrics And Intelligent Laboratory Systems 30*, 133-146.
Tauler, R., Lacorte, S. & Barcelo, D. 1996. Application of multivariate self-modeling curve resolution to the quantitation of trace levels of organophosphorus pesticides in natural waters from interlaboratory studies. *Journal Of Chromatography A 730*, 177-183.
Taylor, J., King, RD., Altmann, T. & Fiehn, O. 2002. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics 18*, S241-S248.
The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature 408*, 796.
Toft, J. 1995. Evolutionary Rank Analysis Applied To Multidetectional Chromatographic Structures. *Chemometrics And Intelligent Laboratory Systems 29*, 189-212.
Tohge, T., Nishiyama, Y., Hirai, MY., Yano, M., Nakajima, J., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, DB., Kitayama, M., Noji, M., Yamazaki, M. & Saito, K. 2005. Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *Plant Journal 42*, 218-235.
Tolstikov, VV. & Fiehn, O. 2002. Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical Biochemistry 301*, 298-307.
Tolstikov, VV., Lommen, A., Nakanishi, K., Tanaka, N. & Fiehn, O. 2003. Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Analytical Chemistry 75*, 6737-6740.
Trygg, J. & Wold, S. 2002. Orthogonal projections to latent structures (O-PLS). *Journal Of Chemometrics 16*, 119-128.
Trygg, J. & Wold, S. 2003. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal Of Chemometrics 17*, 53-64.
Tyler, L., Thomas, SG., Hu, JH., Dill, A., Alonso, JM., Ecker, JR. & Sun, TP. 2004. DELLA proteins and gibberellin-regulated seed germination and floral

development in Arabidopsis. *Plant Physiology 135*, 1008-1019.
Wagner, C., Sefkow, M. & Kopka, J. 2003. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry 62*, 887-900.
Walhout, JS. & Pierce, AE. 1968. *Theoretical Aspects of Thrimetylsilylation*. Ed Pierce Chemical Company. Rockford. pp 60
Ward, JL., Harris, C., Lewis, J. & Beale, MH. 2003. Assessment of H-1 NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis thaliana. *Phytochemistry 62*, 949-957.
Watson, JT. 1997. *Introduction to Mass Spectrometry*. 3nd editionrd editon. Lippincott Williams & Wilkins. New York. 496 pp
Weckwerth, W., Loureiro, ME., Wenzel, K. & Fiehn, O. 2004. Differential metabolic networks unravel the effects of silent plant phenotypes. *Proceedings Of The National Academy Of Sciences Of The United States Of America 101*, 7809-7814.
Verdonk, JC., de Vos, CHR., Verhoeven, HA., Haring, MA., van Tunen, AJ. & Schuurink, RC. 2003. Regulation of floral scent production in petunia revealed by targeted metabolomics. *Phytochemistry 62*, 997-1008.
Veriotti, T. & Sacks, R. 2001. High-speed GC and GC/time-of-flight MS of lemon and lime oil samples. *Analytical Chemistry 73*, 4395-4402.
Wiklund, S., Karlsson, M., Antti, H., Johnels, D., Sjöström, M., Wingsle, G. & Edlund, U. 2005. A new metabonomic strategy for analysing the growth process of the poplar tree. *Plant Biotechnology Journal 3*, 353-362.
Vikram, A., Prithiviraj, B. & Kushalappa, AC. 2004. Use of volatile metabolite profiles to discriminate fungal diseases of Cortland and empire apples. *Journal Of Plant Pathology 86*, 215-225.
Wilson, ID., Nicholson, JK., Castro-Perez, J., Granger, JH., Johnson, KA., Smith, BW. & Plumb, RS. 2005a. High resolution "Ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *Journal Of Proteome Research 4*, 591-598.
Wilson, ID., Plumb, R., Granger, J., Major, H., Williams, R. & Lenz, EA. 2005b. HPLC-MS-based methods for the study of metabonomics. *Journal Of Chromatography B-Analytical Technologies In The Biomedical And Life Sciences 817*, 67-76.
Windig, W. & Guilment, J. 1991. Interactive Self-Modeling Mixture Analysis. *Analytical Chemistry 63*, 1425-1432.
Wolbang, CM., Chandler, PM., Smith, JJ. & Ross, JJ. 2004. Auxin from the developing inflorescence is required for the biosynthesis of active gibberellins in barley stems. *Plant Physiology 134*, 769-776.
Wolbang, CM. & Ross, JJ. 2001. Auxin promotes gibberellin biosynthesis in decapitated tobacco plants. *Planta 214*, 153-157.
Wold, S. 1978. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics 20*, 397-405.
Wold, S., Sjöström, M. & Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics And Intelligent Laboratory Systems 58*, 109-130.
Vorst, O., Vos, CHRd., Lommen, A., Staps, RV., Visser, RGF., Bino, RJ. & Hall, RD. 2005. A non-directed approach to the differential analysis of multiple LC-MS-derived metabolic profiles. *Metabolomics 1*, 169.

Acknowledgements

Jag vill tacka:

Thomas Moritz min handledare och mentor för att du har gett mig frihet att utvecklats under min tid i Umeå. Du har trott och ställt upp på mig undra de här åren. Av dig har jag lärt mig allt som man behöver veta om masspektrometri och att Skellefteå är världens mitt. Du kommer alltid för mig även bli den dåliga ordvitsens okrönta kung. Ett riktigt stort tack till **Krister L** och **Inga-Britt C** för all hjälp på labbet utan er hade jag inte fixat tiden i Umeå. Ni är underbara och gör ett fantastiskt arbete med team Moritz och UPSC. Jag vill även tacka alla nuvarande och alla forna medlemmar i **team Moritz** för gruppmöten och middagar. Även **Göran S** min biträdande handledare som har varit en bra mentor och inspirationskälla fast jag inte ha haft ett enda handledarmöte med dig.

Anders N för att du fick mig att flytta upp till Umeå igen och att du var min inofficiella biträdande handledare. Vi har haft så mycket kul på alla resor och de många kvällarna på Rött. Du är en riktig vän som har betytt och betyder mycket för mig. **Pär J** för vårt fruktbara samarbete. Det har varit väldigt kul att arbeta med dig och jag kan glömma när vi gled runt i en Mustang i Washington. **Pär J**, **Johan T och Michael S** för att ni har orkat lyssna på mina frågor och tossiga idéer under alla dessa år. Ni har varit en stor inspirationskälla och fått mig att förstå mycket om kemometrins underbara värld. Jag vill även tacka **Susanne W**, **Ing-Marie O**, **Henrik A**, **Hasse S** och alla andra som är eller har varit på kemometrin.

Alla ni på kontoret: **Jonathan L**, **Annika J**, **Sara VP**, **Donald T**, den ständiga gästen **Hendrik B** och forna rumskompisar **Anders N**, **Kazou S** och **Avano T**. Vi har nog snackat lite för mycket skit under alla år, men jag kommer alltid att minnas alla roliga upptåg som poesi måndag, mustasch fredag och återkommande pumpatävlingar. Det har varit superkul att dela rum med er. **Annika J** för att du alltid har brytt dig om mig och du kommer med bravur att ta över staffetpinnen inom metabolomics forskningen. **Jonatan L** för din entusiasm, dina dikter och alla galna företags idéer vi har haft. **Sara VP** fortsätt på samma bana så kommer du bli en framgångsrik forskare. Lycka till med kolonilotten!

Hendrik B, instutionens egna solstråle och optimisternas optimist, för att du var min privata biologiska läromästare. Du har fått mig som stadsbo intresserad av biologi, blommor och fåglar. Man kan lita på att du och **Urs F** alltid snackar skit mellan himmel och jord och lite till. **Janne E** du har varit god samtalspartner vad det gäller det mesta som är viktigt här i livet som musik, film och illustrationer. **Maria I** och **Sara A** för att har varit instutionens egna labordningskvinnor, helpline, partyfixare, spexfixare och partypinglor. Ni har förgyllt allas tid på UPSC. **Gertrud L** och **Maria L** för all hjälp och ert idoga arbete att hålla ordning på instutionen. **Stefan L** för all hjälp och kul att ha haft någon att munhuggas med på instutionen. Nu får du klara dig utan Göteborgaren!

The east European mafia: **Mariusz K**, **Karel D**, **Petr T**, **Dana T** and **Vera T** for many laughs and all the crazy Czech dinners. **Mariusz K** for all help and for a great

visit in St Paul. **Camilla V** för att du har lärt mig allt om kromatografins underbara värld. **Erik J** för din entusiasm och samma otröttliga intresse för att optimera med DOE som jag. Alla i mass rummet och inte minst **Sara VP** och **Karin L** för all hjälp med IAA analyserna. **Gun L**, **Janne E** och **Karin L** för foton och illustrationer. Jag vill tacka **Kjell O**, UPSC egen paparazzifotograf, för att ha förgyllt instutionen med roliga festfoton. Tack även för omslags fotot. Vill även tacka alla på jobbet, speciellt **Mattias H**, **Sven E**, **Karin N**, **Lena D**, **Sandra J**, **Daniel E**, **Erling Ö**, **Jenny H**, **Henrik S** och **Jörgen P** för bl a. sällskap i labbet, för diskussioner på luncherna, fikan, fester, Kreta resan och allt annat. Inte att förglömma **Mattias H** för din omtanke.

Anna-Lena S för du inspirerade mig till att börja doktorera, och att du alltid har tid att lyssna och visat intresse för mitt arbete. Tacka till **Peter v** för musiktipsen och att du fixade lägenhet till mig. Även alla andra på Arbetslivsinstitutet. **Anders Ö** och **Krister Å** för att ni har varit inspirationskällor och bra bollplank. **Hans W** för ditt engagemang för mig och din generositet under tiden på Eka Chemicals.

Calle V, min trogna vän, ”bror” och vapendragare. Du har entusiasmerat mig till mycket och utan våra otaliga telefonsamtal hade jag haft en tråkig tid i Umeå. **Janne G** för att i unga år var den bästa sparringpartnern. Kompisarna **Micke J**, **Malin L**, **Olof B, Agneta E** och **Jenny R**.

Kristina L för att du fick mig att inse att jag hade en talang och att dyslexin inte skall begränsa mig. **Marja L** för all hjälp och insikt. Och min älskade familj som har stöttat mig och varit intresserat av mitt arbete. Min pappa **Benny G** som alltid har varit och kommer vara min ständiga sparringpartner i allt. Du har lärt mig att inget är omöjligt och att man skall hålla drömmen vid liv. Mamma **Barbro G** som har brytt sig, oroat sig och glatt sig med mig. Min underbara syster **Jenny G** och hennes **Göran P** som alltid har varit intresserad vad jag har gjort fastän ni inte har riktigt förstått vad jag forskat om. Och inte minst min syster dotter **Julia P** för att du är så fantastisk.

Och min kära **Elisabeth B** för att du alltid bryr dig om mig och tar hand om mig Du får mig att skratta och uppskatta livets små glädjeämnen. Utan dig hade jag inte klarat mitt sista halvår på UPSC. Resan har börjat. **Familjen Burström** för alla utflykter och er gästfrihet och **Bettans kompisar** för allt bus.

Och till sist alla er som har hjälp mig med min dyslexi och korrekturläst mina texter: **Johan T**, **John B**, **Annika J**, **Sara VP**, **Krister L**, **Jenny H**, **Janne E**, **Anders N**, **Jonatan L**, **Mariusz K**, **Susanne W** och inte minst min handledare **Thomas M**. Ni har varit ovärderliga för mig.

Arbetet genomfördes med hjälp av finansiering från **EU:s strukturfonder Mål 1 Norra Norrland** under ledning av ”**Fibernätverket”**. Jag vill även tacka för finansiering från **Wallenberg Consortium North (WCN)** och **Kempe stiftelsen**.