Phytochemical Analysis

Received: 19 August 2009; Revised: 5 October 2009; Accepted: 6 October 2009 Published online in Wiley Interscience: 10 November 2009

(www.interscience.wiley.com) DOI 10.1002/pca.1185

Application of NMR in Plant Metabolomics: Techniques, Problems and Prospects†

Jan Schripsema

ABSTRACT: The present state-of-the-art of NMR in plant metabolomics is reviewed. Attention is paid to the different practical aspects of the application of NMR. The sample preparation, the measurement of the spectrum, quantitative aspects and data analysis are discussed. Each stage has its specific problems, which are pointed out and recommendations are made. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: metabolomics; plants; review; nuclear magnetic resonance spectroscopy

Introduction

Metabolomics is by now becoming a well-known technique for the study of all types of organisms, and complements the data obtained by the other 'omics': genomics, transcriptomics and proteomics. However, what does the term metabolomics stand for and how does it relate to other terms that have emerged, such as metabolic profiling, metabolic fingerprint, or metabonomics? Dettmer *et al*. (2007) presented the following list of definitions, to which the definition of metabonomics, presented by Nicholson and Lindon (2008) was added:

- *Metabolites*—small molecules that participate in general metabolic reactions and that are required for the maintenance, growth and normal function of a cell.
- *Metabolome*—the complete set of metabolites in an organism.
- *Metabolomics*—identification and quantification of all metabolites in a biological system.
- *Metabonomics*—measurement of the global, dynamic metabolic response of living systems to biological stimuli or genetic manipulation (Nicholson and Lindon, 2008).
- *Metabolic profiling*—quantitative analysis of set of metabolites in a selected biochemical pathway or a specific class of compounds. This includes target analysis, the analysis of a very limited number of metabolites, e.g. single analytes as precursors or products of biochemical reactions.
- *Metabolic fingerprinting*—unbiased, global screening approach to classify samples based on metabolite patterns or 'fingerprints' that change in response to disease, environmental or genetic perturbations with the ultimate goal to identify discriminating metabolites.
- *Metabolic footprinting*—fingerprinting analysis of extracellular metabolites in cell culture medium as a reflection of metabolite excretion or uptake by cells.

Depending on the paper and the organism being studied, slight variations in these definitions can be found, but it is clear that the metabolome involves all metabolites in an organism and that the intention of metabolomics is to identify and quantify all those metabolites. However, many metabolomic studies do not go further than the metabolic fingerprinting stage: producing metabolite patterns by high-throughput analytical methods (MS or NMR) and searching for discriminating factors by chemometric methods. Next to fingerprinting, the term footprinting is also used, but this applies especially to cell culture systems (Allen *et al*., 2003; Mashego *et al*., 2007). In metabolic profiling specific groups or categories of metabolites are defined for precise quantification, but this yields only a limited view of the total metabolome (Scalbert *et al*., 2009). In metabonomics the aim is to measure the global, dynamic metabolic response of living systems to biological stimuli or genetic manipulation (Nicholson and Lindon, 2008). The focus is on understanding the systemic change through time in complex multicellular systems. In practice, there is no real difference between metabonomics and metabolomics, but the term metabonomics is traditionally more used in biomedical research to describe the fingerprinting of biochemical perturbations caused by disease, drugs, and toxins (Goodacre, 2007; Davies, 2009).

History of Metabolomics

In 1998, the term metabolome was introduced in a paper from Oliver *et al*. about the yeast genome, based on the analogy with the terms genome, proteome and transcriptome, which had been introduced already. From the year 2000 the term metabolomics came into use in publications, as can be seen from Fig. 1. This figure is based on the number of papers found with the respective terms in literature, as compiled by ISI Web of Science (Thomson Scientific). The term genome was the first to be introduced, and from 1988 the term genomics came in use. Proteomics

* Correspondence to: J. Schripsema, Grupo Metabolômica, Laboratório de Ciências Quimicas, Universidade Estadual do Norte Fluminense, Av. Alberto Lamego, 2000, 28015-620 Campos dos Goytacazes, RJ, Brazil E-mail: jan@uenf.br

† This article is published in *Phytochemical Analysis* as a special issue on Metabolomics in Plant and Herbal Medicine Research, edited by Young Hae Choi, Hye Kyong Kim and Robert Verpoorte, all from Leiden University in the Netherlands.

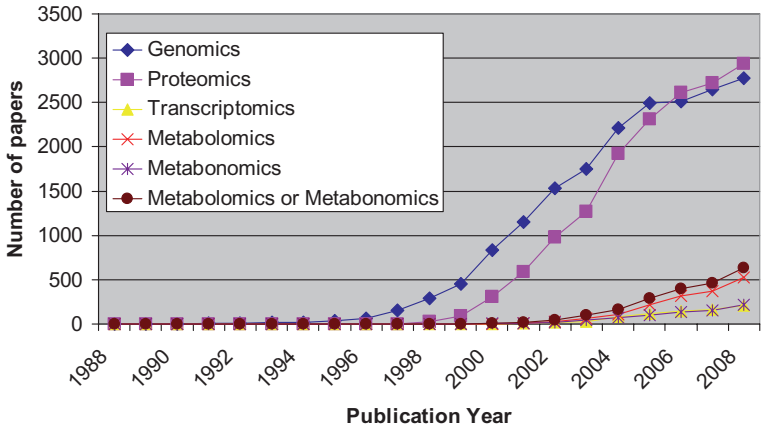

Figure 1. Number of papers found with the respective terms in literature, as compiled by ISI Web of Science (Thomson Scientific).

followed only in 1997 and transcriptomics in 1999 (Fig. 1). In Fig. 1 it can be seen that the term genomics only came into common use rather long after it had been introduced, but the number of papers rapidly increased since 1996 to about 2500–3000 papers a year over recent years. Proteomics research rapidly grew after 1998 and the number of papers surpassed the number of papers about genomics since 2006, but like the number of papers on genomics, it seems to have reached its plateau, with about 3000 papers a year. The number of papers about metabolomics (and metabonomics) shows a slower growth and the number of papers reached more than 600 in 2008. The slow growth might be due to the fact that metabolomics has no well-defined methods of analysis (see below), like genomics and proteomics. Nevertheless, it seems probable that growth will continue and might reach similar levels to those of proteomics and genomics. When the term metabolomics was introduced, it was not a new process or new concept, but it was more a consequence of the development of the techniques which form the basis of metabolomics. In fact the metabolite profiling which is the basis of metabolomics already had been performed a long time before. One can think of the TLC profiles of medicinal plant extracts, or the GC chromatograms of essential oils. The increased resolution in NMR spectra by higher field magnets and the wider application of two-dimensional techniques permitted the identification of metabolites in crude extracts or samples. Already from the mid-1980s unmodified biological fluids were being studied by NMR (Nicholson and Lindon, 2008), and in the early 1990s metabolic fingerprinting of plant materials came into use (Schripsema and Verpoorte, 1991; Schripsema *et al*., 1991). Also the development in chromatographic techniques, such as capillary columns for GC and small particle HPLC columns, permitted a wider application of these techniques. GC was successfully applied for the separation of relatively polar underivatised substances, such as for instance monoterpenoid indole alkaloids (Dagnino *et al*., 1991). By derivatisation the range of compounds which could be analysed was further amplified. By coupling to mass spectrometry also the direct identification of the compounds was possible. This then led to the availability of techniques for comprehensive profiling of samples (e.g. Fiehn *et al*., 2000). Another important contribution to metabolomics came from the advances in chemometrics, which permitted the evaluation of large data sets and distilling significant changes in relation to specific parameters.

Techniques for Metabolomics

Metabolic fingerprinting can be done with the most diverse techniques, spectroscopic like UV, IR (Goodacre, 2005; Defernez and Wilson, 1997), MS and NMR or chromatographic like GC or HPLC, or using hyphenated techniques, like GC-MS or LC-NMR. Several reviews give a quite detailed report about the various technologies used for metabolomics, e.g. Schauer and Fernie (2006) and Hall (2006). Nowadays in metabolomics research, where issues of quantification and identification are important, NMR and mass spectrometry are the principle detection techniques to be used, each of them having specific advantages and disadvantages.

Within mass spectrometry many different techniques can be distinguished, varying in the method of introduction of the sample (direct, GC, LC), in the method of ionisation (electron impact, MALDI, electrospray) and in the method of detection (time-of-flight, FT-ICR). The most commonly used mass spectrometry platforms are:

- GC-MS—the high chromatographic resolving power of gas chromatography is combined with electron impact mass spectrometry. Electron impact ionisation provides for each compound a mass spectrum, which through the fragmentations provides much information about its identity. Because EI is the most traditional way of ionisation, large databases exist and compounds can be rapidly identified. A major disadvantage of GC-MS is the fact that compounds need to be volatile to be analysed. Furthermore they should be stable during the analysis in which high temperatures are used. By derivatisation many normally not volatile compounds can be converted in volatile adducts, e.g. by silylation, acetylation or methylation. In this way, for example, sugars and amino acids can be analysed.
- LC-MS—liquid chromatography is combined with mass spectrometry. In this type of analysis all types of compounds can be separated by LC and subsequently in-line analysed by MS, usually electrospray-MS. With modern columns (e.g. UPLC) high resolutions can be obtained. The electrospray mass spectra are, however, less informative then the EI mass spectra. Also large databases are not (yet) available. Another problem in electrospray mass spectrometry is the difficulty of ionisation of many compounds. It is however possible to work in positive or negative ion mode.

- FT-ICR-MS—in Fourier transform ion cyclotron resonance MS, the mass spectrometer has a very high resolution. This gives complete resolution of compounds based on the exact masses. However, isomers cannot be resolved. For metabolomics the chromatography step, which might prevent certain compounds to arrive to the mass spectrometer, is often omitted. In Huang-qin extracts it has been reported that more than 2000 compounds were separated (Murch *et al*., 2004).
- CE-MS—capillary electrophoresis coupled to MS. The coupling of CE to MS is rather new, but promising results have been obtained (Monton and Soga, 2007).

Mass spectrometry has a very high sensitivity and large dynamic range. In metabolomic research easily hundreds of compounds can be detected. In NMR the sensitivity is much less, but the structural information content, reproducibility and quantitative aspects are superior to mass spectrometry. Furthermore the preparation of the sample is simpler and the analysis more rapid. This makes NMR the ideal tool for broad-range profiling of abundant metabolites and for metabolite fingerprinting of extensive sample collections (Lommen *et al*., 1998; Dixon *et al*., 2006). The number of compounds which can be detected in a single analysis is however limited from one to several dozens (Krishnan *et al*., 2005, Martin *et al*., 2007). In mass spectrometry these numbers are much larger. A metabolomic study of tomato by FTICR-MS led to the recognition of 869 metabolites. By comparison with public databases, it was suggested that 494 of these metabolites are novel (Iijima *et al*., 2008). However, it should be considered that in mass spectrometry a single MH^+ signal indicates the presence of a compound with a specific molecular formula, but no structural information is obtained and many isomers might exist. In ^1H NMR each signal corresponds to a specific hydrogen atom within a molecule and within a spectrum for every hydrogen atom of a molecule a signal is found. Isochronic signals of different compounds occur. Therefore, for identification all signals of a compound should be considered. When compounds are identified using single signals, by a simple database search, mistakes are possible.

Experimental Procedures for NMR in Metabolomics

Metabolomics is focussed on the complete analysis of all metabolites inside a certain sample. One of the major problems resides in the problem to obtain a sample which permits the complete analysis of all metabolites inside it, which is virtually impossible. If this sample is treated is some way, by extraction, separation or derivatisation, information will be lost. In an extraction only the soluble components are obtained and insoluble or partly soluble compounds are lost. When a separation is performed, certain parts of the sample are also lost. In derivatisation new compounds are obtained, but the structures are changed.

Furthermore all the metabolites should be analysed, and in metabolomics one is required to analyse a wide variety of different compounds, which differ in many aspects:

- chemical nature, e.g. alkanes, carboxylic acids, amines, esters, peptides;
- solubility, ranging from water soluble, such as sugars, to oil soluble, such as lipids;
- concentration, ranging from very high concentrations, such as sugars (e.g. 50% of a sample) to minute concentrations for signal compounds which occur in picomolar quantities (in the range of pg to ng quantities per ml).

This is very different in relation to the other omics technologies: in genomics one deals with the DNA, which is a specific chemical entity consisting of only four basic units, the nucleotides. In proteomics one deals with proteins, which are built from 20 amino acids. In transcriptomics one deals with mRNA, also built with the same basic four nucleotides. If one wants to analyse the complete metabolome, multiple analysis will be required, focussed on specific parts. These parts can be determined by the abovementioned factors: chemical nature, solubility and concentration. It will be important then to know how many metabolites are present, but only rough estimates of this have been made. In *Arabidopsis* plants there are expected to be about 5000 chemical entities, both primary and secondary metabolites (Bino *et al*., 2004). Furthermore, at the moment about 200000 natural compounds are known (Dixon and Strack, 2003), but certainly many more await discovery. What makes plant metabolomics much more complex than metabolomics of humans and animals is this enormous diversity of chemical structures, especially within the so-called secondary metabolites. Every species has a specific set of metabolites and this has been exploited traditionally for chemosystematics. The quantities of these secondary metabolites also show large variations and in many cases their concentrations surpass those of common primary metabolites. In NMR spectra from plant tissues commonly secondary metabolites are directly observed (Schripsema *et al*., 2007). Even in cell cultures, which generally have lower levels of secondary metabolites, they can be observed in crude extracts.

In NMR-based metabolomics data collection and spectral processing are also important to ensure that, for example, replicate samples provide identical NMR fingerprints. In practice small differences in lineshape and chemical shift will be observed. The differences in lineshape can be minimised by using exactly the same sample volume in identical NMR tubes and by optimising the magnetic field homogeneity before data acquisition (Krishnan *et al*., 2005). To compensate for diffences in linewidth, the line-broadening parameter can be varied during processing (Lommen *et al*., 1998). To minimise the misalignment of NMR signals, there should be a stringent control of sample preparation, especially to avoid differences in pH or ionic strength (Defernez and Colquhoun, 2003).

Sample Preparation

Most laboratories use their own specific method of sample preparation. In metabolomic studies it is important that the reproducibility of the procedure is as best as possible (Defernez and Colquhoun, 2003). All sources of variation should be minimised. That is important in the sample selection, preparation and during measurement (Maher *et al*., 2007). When leaves from a plant are collected, one should consider among others the position of the leaf, its age, exposure to sunlight and rain as well as the time of collection and the weather. All these factors potentially cause variation. After collection the sample treatment is important. First of all storage: at room temperature or frozen. Then the processing of the sample should be considered—artefacts might arise from interaction with solvents or by residual enzymatic activity.

To analyse the metabolome as completely as reasonably possible, a single solvent will not be sufficient (Verpoorte *et al*., 2007). A reasonable coverage of all metabolites can be obtained with two solvents: one apolar extract (e.g. with chloroform) and a polar extract (e.g. with water or water–methanol).

A problem with water extracts is often the high quantity of sugars. The signals of the sugars obscure other signals and limit the dynamic range of the spectrum. Furthermore, due to the sugars, pre-concentration of the sample will be difficult. A similar problem can occur with chloroform extracts. Lipids are generally the major components of the extract.

When there are problems with superimposed signals, better results might be obtained by the use of two-dimensional NMR, e.g. *J*-resolved NMR (Viant, 2003; Liang *et al*., 2006a, b; Widarto *et al*., 2006), HSQC (Fan *et al*., 2001), TOCSY or HMBC (Widarto *et al*., 2006), or further fractionation of the samples should be undertaken.

Protein depletion of samples is generally not considered to be essential, but it should be considered that in water extracts the catalytically active proteins might interfere in the spectrum. For example, sucrose in papaya extracts is rapidly converted by invertase activity into glucose and fructose (Schripsema *et al*., 2009). The removal of proteins might be considered; however, for the most rapid reactions before or during the extraction the proteins should be deactivated. For serum samples a number of deproteinisation procedures were investigated by Daykin *et al*. (2002) and later by Tiziani *et al*. (2008). They found that the best results were achieved by ultrafiltration, which removes proteins quantitatively, yields good signal-to-noise, and is superior in reproducibility. If in the sample preparation procedure the sample is dried and redissolved in deuterated solvent, one should consider that volatile components are completely or partially lost, e.g. ethanol (Tiziani *et al*., 2008) or salicylic acid (Verpoorte *et al*., 2008). For extraction of biological material one has to select a solvent or solvent mixture from the many options available. For this selection a number of parameters should be considered (Verpoorte *et al*., 2008):

- *Polarity and selectivity of the solvent*—these determine the solubility of the individual components in the solvent. The solvent strength of solvents as indicated in Table 1 gives a reasonable impression of the type of compounds which might be extracted. However, acid/base behaviour should also be considered.
- *Boiling point* is important if solvents need to be evaporated, which might lead to thermal decomposition or loss of compounds by evaporation. Water as a solvent, despite the relatively high boiling point, has the advantage that it can easily be removed by lyophilisation.
- *Toxicity and environmental considerations*—for example, benzene should be avoided due to its carcinogenicity and can in extractions easily be substituted with toluene. Also chloroform can in many cases be substituted with dichloromethane, because the latter is less toxic.
- *Possible contaminations in solvents* that may interfere with the analysis, yielding unwanted signals in the spectra, e.g. antioxidants such as butylated hydroxytoluene, 2,6-di-*tert*-butylphenol and propylgallate (in ethers, chloroform).
- *Possible contaminations that may cause artefact formation*, e.g. peroxides (in ethers), dichlorocarbene (in chloroform), ethylchloroformate (in chloroform), phosgene (in chloroform), dichlorobromomethane (in dichloromethane; chloroform), ethanol (in chloroform 1–2% of ethanol as stabiliser; diethyl ether; ethyl acetate), formaldehyde (in chloroform; ethyl acetate) and acetaldehyde (in chloroform; ethyl acetate). The artefact formation has been discussed more extensively by Verpoorte *et al*. (2008).

Table 1. Eluotropic series of solvents listed according to their eluting power on alumina adsorbent. The empirical solvent strength data were reported by Snyder (1968)

Solvent	Solvent strength
Water	>>1
Methanol	0.95
Ethanol	0.88
2-Propanol	0.82
Dimethyl sulfoxide	0.75
Pyridine	0.71
Acetonitrile	0.65
Ethyl acetate	0.58
Tetrahydrofuran	0.57
Acetone	0.56
Dichloromethane	0.42
Chloroform	0.40
Diethyl ether	0.38
Benzene	0.32
Toluene	0.29
Hexane	0.01

Material to be extracted can be fresh, lyophilised or dried in a different way (air-dried). Within the different experimental procedures for the NMR sample preparation which have been used for metabolomics research, two approaches can be distinguished:

1. *Direct extraction with deuterated solvent*. In this case, after a centrifugation step, the supernatant is directly used for the NMR analysis.
2. *Previous extraction of the biological material by non-deuterated solvents*. The resultant solution is dried under vacuum and subsequently re-dissolved in deuterated solvent. After this re-dissolution generally a new centrifugation step is necessary before the liquid is used for NMR analysis. For the direct extraction with deuterated solvent, the most common approach is the extraction of the polar metabolites with mixtures of methanol and buffered water. In many cases chloroform is also added, which yields a separate phase concentrating the apolar constituents. Variations are found in the relative proportions of methanol and water, and in the pH of the water phase. Several studies reported the use of 50% methanol-d_4–50% D_2O (KH_2PO_4 buffer pH 6), together with an equal volume of $CDCl_3$ (Hendrawati *et al*., 2006; Widarto *et al*., 2006; Liang *et al*., 2006b; Sanchez-Sampedro *et al*., 2007; Simoh *et al*., 2009; Leiss *et al*. 2009). In a study on potato, tomato and tea 70% methanol-d_4–30% phosphate buffer was used (Defernez and Colquhoun, 2003).

When non-deuterated solvents are used for the extraction, generally larger quantities of solvent and biological material are used, such that after drying the extract upon re-dissolving, a higher concentration of the metabolites can be obtained (Flores-Sanchez *et al*., 2009; Schripsema *et al*., 1991). It should be noted however, that generally not everything re-dissolves and an additional centrifugation step will be necessary.

NMR Measurement

After preparation of the samples they should be submitted to NMR. Liquid samples in non-deuterated solvents can be submitted to NMR, but in that case a little deuterated solvent should be added to provide the lock signal for NMR and during the measurement solvent suppression should be applied. When the samples are dry they should be dissolved in deuterated solvents. The commonly available deuterated solvents are D_2O, methanol-d_4, DMSO-d_6, acetone-d_6, chloroform-d and benzene-d_6. First of all the choice of solvent will be determined by the nature of the sample and its preparation. For polar extracts usually D_2O, methanol-d_4 or a mixture of the two is used. For apolar extracts usually deuterated chloroform is used. However, a number of factors should be considered in the choice.

- *Solubility of the extract*—if, for example, an alcoholic extract has been made and for NMR measurement D_2O is chosen the more apolar components of the extract might not be dissolved.
- *NMR spectra reported in literature* are (nearly) always recorded in one of the mentioned deuterated solvents, rarely in mixtures. If a spectrum is obtained in another solvent or solvent mixture, the chemical shifts will be different.
- *The price of the solvents*—D_2O and deuterated chloroform are rather cheap, but deuterated methanol is much more expensive.
- If *aqueous samples* are used it should be decided whether or not pH control is necessary. Variation in pH between samples causes changes in the chemical shifts of certain signals. In particular, the signals from citric acid, which is a very common constituent of plant extracts, are sensitive to pH differences. Buffers which have been used are most commonly phosphate buffers of pH 6. Also, phosphate buffer, pH 7 (Aranibar *et al*., 2006), and oxalate buffer, pH 4 (Pereira *et al*., 2005; Son *et al*., 2009) have been reported.
- Another point to be considered is the inclusion of an *internal standard*. Usually a standard, TMS or a derivative, is used as chemical shift reference. If the concentration of the internal standard is exactly known, it might also be used for quantitative purposes.

When a solvent suppression sequence is used for NMR acquisition and quantitative analysis of the signals is required, some care should be taken with the selection of the sequence and experimental details. This is described by Saude *et al*. (2006). The methods of registration and processing of the NMR spectrum all have effects on the final spectrum. Special care should be taken with the apparatus adjustments, such as shimming and temperature, with the acquisition parameters, such as the acquisition and relaxation times, spectral width and number of data-points for the measurement. Also, during processing care should be taken with the phasing, line broadening and zero filling (Defernez and Colquhoun, 2003). However, it is most important to avoid variation within series of samples.

Quantitative Aspects of ^{1}H NMR

In principle ^{1}H NMR provides a reliable profile of each sample, and quantities are reflected in the integrals of the individual signals of the spectrum. This makes ^{1}H NMR a unique technique, which enables the quantification of compounds in relation to any other compound in the spectrum (e.g. Dagnino and Schripsema, 2005). In mass spectrometry quantification is much more difficult and the intensity of signals depends on many factors, instrument-related and related to other compounds in the sample, e.g. through ion-suppression.

Quantitative measurements by ^{1}H NMR need some specific precautions, aside from the obvious need for an adequate signal to-noise ratio in the spectrum (Claridge, 1999). These are the avoidance of differential saturation effects and the need to characterise the NMR resonance line-shape properly (using at least 4 points/Hz), so it is beneficial to use the minimum spectral width compatible with the sample and to adjust the acquisition times accordingly. Results can be enhanced further during data processing: to ensure that the data are sufficiently well digitised in addition to improving the signal-to-noise ratio, an exponential function, slightly broadening the lines, should be applied (0.3 Hz). Furthermore, zero filling with a factor of 4 can be applied. To completely avoid the differential saturation effects, the spins should fully relax between pulses, demanding recycle times of at least 5 times the T_1 (longitudinal relaxation or spin-lattice relaxation time) of the slowest relaxing nuclei. In most published metabolomics experiments the recycle time which has been used (in the order of 5–10 s), was not sufficient to achieve a complete relaxation of all nuclei, considering that formic acid, a commonly present slow relaxing compound, has a T_1 of 8.0 s. Also TMSP, commonly used as internal standard in D_2O solutions, has a rather long T_1 of 3.2 s (Schripsema, 2008). For a table of T_1 values of common metabolites see Weljie *et al*. (2006). To obtain the best accuracy for the quantitations, instead of increasing the recycle time, a correction can be made for the differences in T_1 using a mathematical correction factor for quantitation (Saude *et al*., 2006). Of course, it depends on the purpose of the measurements if this is necessary. If a very accurate quantitation (+1%) of a single compound is required, the correction factor can be used, and it would also be advisable to make a calibration curve of that compound. Special care should also be taken with the start and finish of the integral: for a NMR signal (Lorentzian line shape), the integral should ideally cover 20 times the half-height line-width on each side of the peak if it is to include 99% of it (10–20 Hz each side) (Claridge, 1999). For the absolute quantification of metabolites an exactly known quantity of internal standard should be present in the NMR sample. Any compound with signals not overlapping with other signals from the sample might be suitable. Compounds which have been used for this purpose in chloroform solutions are hexamethyldisilane (Choi *et al*., 2004; Flores-Sanchez *et al*., 2009; Leiss *et al*., 2009) and pyridine (Dagnino and Schripsema, 2005). In aqueous solutions phloroglucinol (Choi *et al*. 2003), sodium trimethylsilyl propionate (TMSP) (Choi *et al*., 2004, 2006; Flores-Sanchez *et al*., 2009, Leiss *et al*., 2009) and 4,4-dimethyl-4-silapentane-1-sulfonic acid (Weljie *et al*., 2006) have been used.

An alternative for the use of an internal standard for quantification is the use of an electronically generated reference signal (Akoka *et al*., 1999). This is known as the ERETIC method (electronic reference to access *in vivo* concentrations). The advantages are that no internal standard needs to be added and that the reference signal frequency can be freely chosen to fall within a transparent region of the spectrum (Akoka *et al*., 1999). However, a calibration of the ERETIC peak with a standard solution will be necessary. Maybe in the near future this method will be routinely available on NMR spectrometers.

Data Processing

After the NMR spectra have been obtained, they should be processed to extract the data. Basically, three methods to extract the data can be distinguished (Scalbert *et al*., 2009): Binning, peakpicking or deconvolution. One of these techniques is necessary, because the crude NMR data show artefacts due to physicochemical differences (Torgrip *et al*., 2006). Major data artefacts are considered: peak shifts, peak shape distortions (shim problems) and unsuccessful phasing. Peak shifts are common and can be due to pH variation between samples, variation in concentrations, interactions between sample components or temperature variations. Generally these shifts are very small (in the order of 0.01 ppm), but they interfere in the direct evaluation of spectra.

Binning or bucketing is the most common approach in metabolomic research (Spraul *et al*., 1994). The signal in fixed chemical shift regions, generally with a fixed width of about 0.04 ppm, is averaged. In general, this reduces the 16K datapoints, containing information on hundreds of peaks, to a mere 250 datapoints. The result of this is of course a tremendous loss of resolution. However, the comparison of spectra is facilitated and by subsequent chemometric analysis, areas containing information about, for example, biomarkers can be revealed.

In the peakpicking approach, due to the beforementioned physicochemical differences, a subsequent alignment procedure will be necessary. The partial linear fit method was described by Vogels *et al*. (1996). Forshed *et al*. (2005) investigated two dedicated peak alignment methods and found that both produced better results than the bucketing approach.

The best approach, however, is the deconvolution of the NMR spectra into the individual compound spectra. This approach has been described by Weljie *et al*. (2006) and the name 'targeted profiling' was proposed. The NMR spectrum is mathematically modelled from pure compound spectra, based on the interrogation of a database to identify and quantify the metabolites in the mixture. This approach was tested for synthetic and normal urine samples and yielded more reliable data compared with spectral binning (Weljie *et al*. 2006). The targeted profiling would be of great value for the characterisation of compounds at low concentrations, whose signals are overlapped. In binning the high intensity signals would dominate the analysis and overwhelm the maybe more important changes of low-concentration metabolites.

The extraction of the metabolite information as done in the deconvolution of the spectra and submission of that information to subsequent PCA analysis would be preferred, but often it is not feasible (yet) or too elaborate, because especially in plant extracts many unknown (secondary) metabolites might be present. The new method statistical total correlation spectroscopy (STOCSY) offers a further step forward (Cloarec *et al*., 2005; Lindon *et al*., 2006; Sands *et al*., 2009). This method permits the identification of multiple NMR peaks from the same molecule in a complex mixture. It takes advantage of the multi-colinearity of the intensity variables in a set of spectra to generate a pseudo-two-dimensional NMR spectrum that displays the correlation among the intensities of the various peaks across the whole sample.

In metabolomics the differences between datasets are important. To find those differences in the data in an objective way, they are submitted to multivariate analysis, which can be a principal component analysis (PCA) or a partial least squares differential analysis (PLS-DA), in which the data are correlated to an independent variable. An extensive review about pattern recognition methods and their applications in biomedical research has been published by Lindon *et al*. (2001). PLS regression in chemometrics was reviewed by Wold *et al*. (2001).

A problem in this approach might be the over-fitting of data (Broadhurst and Kell, 2006; Rubingh *et al*., 2006; Westerhuis *et al*., 2008). This over-fitting is considered the greatest multivariate analysis problem (Scalbert *et al*., 2009). It is the consequence of the ability of PLS-DA to discriminate even random data sets into two groups with perfect separation between the arbitrary classes in the PLS score plot (Westerhuis *et al*., 2008). To avoid erroneous conclusions an adequate cross model validation is required. However, it should be remembered that the lower the number of subjects compared with the number of variables, the less the outcome of validation tools such as cross-validation, jack-knifing and permutation tests can be trusted (Rubingh *et al*., 2006). The gold standard would be a biological replication in a blind, new dataset (Scalbert *et al*., 2009).

When significant differences between the datasets are found, it is important to determine what is the physiological background of the differences. First of all the metabolites responsible for the differences should be identified. However, this is the most difficult and time-consuming part of a metabolomics study, especially when plant material is being investigated. Many studies exist in which the identification process has been ignored, but without formal compound identification, the discovery of any metabolically interesting patterns or clusters (via PCA or PLS-DA) is largely meaningless (Scalbert *et al*., 2009).

In plant material the identification of metabolites is expected to be much more difficult, because the variety of structures in plants is much bigger than in animals or humans. It has been estimated that the plant kingdom contains more than 200000 metabolites (Dixon and Strack, 2003). For a single species several thousand of metabolites are expected. For *Arabidopsis* an estimate of ca. 5000 has been made (Bino *et al*., 2004).

The correct identification of a compound can be extremely difficult, and much care should be taken when identifying compounds in a metabolomic study. For the definite identification of molecules generally a series of data is necessary, which can be obtained with different techniques, besides NMR, e.g. mass spectrometry, UV–vis and IR spectroscopy. All data should fit with the data obtained from literature or through a database, if it is a known compound. In rare cases a single signal in the ^{1}H-NMR spectrum can be taken as the proof for the identity. For example, when the presence of a certain metabolite is highly probable, or if it has been isolated and identified before from the same material. Anyway, a standard addition of pure reference compound to verify if the signal exactly coincides is advisable. Generally compounds contain more signals in the ^{1}H-NMR spectrum. When analysing mixtures some or all of those signals might be superimposed with signals from other compounds. However, anyway when a compound is present, all signals in the spectrum of the pure compound should be present in the spectrum of the mixture. When the presence of a compound is suspected, this should be confirmed. Various two-dimensional NMR methods can be used for this purpose. The most important of these are:

- 2D-COSY—shows correlations between scalar coupled hydrogen atoms;
- 2D-HSQC—shows correlations between hydrogen atoms and the carbon atom to which they are attached;
- 2D-HMBC—shows correlations between hydrogen atoms and carbon atoms, generally two or three bonds away;

- 2D-TOCSY—shows correlations between all hydrogen atoms of a spin system;
- *J*-resolved—chemical shift and coupling information are displayed in separate dimensions.

The above-mentioned STOCSY is also important, but has not yet been introduced as a standard technique in NMR laboratories.

Through the heteronuclear 2D correlated experiments, HSQC and HMBC, additional information about the chemical shifts of the carbon atoms of the molecule can be obtained. With these techniques generally sufficient information can be obtained for the tentative identification of most primary metabolites and known secondary metabolites if they can be observed in the 1D spectrum. When novel or unknown secondary metabolites are detected, it will generally be necessary to proceed to the isolation of those compounds to be able to identify them.

Conclusions and Future Prospects

Metabolomics has gone through a great development in its few years of existence, but many more applications are awaiting. Compared with metabolomics in humans or animals, the great diversity of secondary metabolites in plants makes it much more difficult to arrive at a routine procedure for sample preparation. For every experiment some optimisation of the sample preparation procedure will be required. In NMR-based metabolomics, or a single polar extract is used, prepared with water or alcohol or a mixture of the two, or a polar extract in combination with an apolar extract, the latter generally made with chloroform.

To overcome the limitation of 1D NMR that only a limited number of metabolites can be observed, increased use of 2D NMR is expected, especially *J*-resolved, TOCSY and HSQC spectra. This is also useful to reveal hidden signals and important to identify the metabolites, permitting eventually an automatic processing of spectra, such as has been reported by the use of the NMR-based metabolomics tool 'MetaboMiner' (Xia *et al*., 2008). However, the most serious obstacle to a more complete analysis of the metabolome may be the difficulty of detecting minor components in the presence of much larger signals (Krishnan *et al*., 2005). To overcome this obstacle further separation of the extracts will be necessary. In particular, inline LC combined with SPE seems very promising (Exarchou *et al*., 2003; Sprogoe *et al*., 2008). Peaks of the LC are collected on SPE cartridges and can be further concentrated by repeated LC. The collected material is washed off with deuterated solvent and high-quality spectra can be obtained. Combination with other techniques is also possible, e.g. UV or MS. In the data processing of metabolomic data, more emphasis will be necessary on the validation of results obtained by multivariate analysis, considering the serious risk of overfitting of data (Broadhurst and Kell 2006; Rubingh *et al*. 2006; Westerhuis *et al*. 2008).

Finally, the identification of marker metabolites will be a major task. Adequate identification of metabolites will be necessary, combining all essential information to be obtained not only by NMR but also by MS, UV, IR, circular dichroism or polarimetry, dependent on the type of molecule.

References

Akoka S, Barantin L, Trierweiler M. 1999. Concentration measurement by proton NMR using the ERETIC method. *Anal Chem* **71**: 2554–2557.

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. *Nat Biotechnol* **21**: 692–696.

Aranibar N, Ott K-H, Roongta V, Mueller L. 2006. Metabolomic analysis using optimized NMR and statistical methods. *Anal Biochem* **355**: 62–70.

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 Plant Sci* **9**: 418–425.

Broadhurst D, Kell D. 2006. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2**: 171–196.

Choi YH, Choi HK, Hazekamp A, Bermejo P, Schilder Y, Erkelens C, Verpoorte R. 2003. Quantitative analyses of bilobalide and ginkgolides from *Ginkgo biloba* leaves and ginkgo products using ^{1}H-NMR. *Chem Pharm Bull* **51**: 158–161.

Choi YH, Kim HK, Hazekamp A, Erkelens C, Lefeber AWM, Verpoorte R. 2004. Metabolomic differentiation of *Cannabis sativa* cultivars using ^{1}H-NMR spectroscopy and principal component analyses. *J Nat Prod* **67**: 953–957.

Choi YH, Kim HK, Linthorst HJM, Hollander JG, Lefeber AWM, Erkelens C, Nuzillard J-M, Verpoorte R. 2006. NMR metabolomics to revisit the tobacco mosaic virus infection in *Nicotiana tabacum* leaves. *J Nat Prod* **69**: 742–748.

Claridge TDW. 1999. *High-Resolution NMR Techniques in Organic Chemistry*. Pergamon Press: New York.

Cloarec O, Dumas ME, Craig A, Barton RH, Trygg J, Hudson J, Blancher C, Gauguier D, Lindon JC, Holmes E, Nicholson J. 2005. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. *Anal Chem* **77**: 1282–1289.

Dagnino D, Schripsema J. 2005. ^{1}H NMR quantification in very dilute toxin solutions: application to anatoxin-a analysis. *Toxicon* **46**: 236–240.

Dagnino D, Schripsema J, Peltenburg A, Verpoorte R, Teunis K. 1991. Capillary gas chromatographic analysis of indole alkaloids: Investigation of the indole alkaloids present in *Tabernaemontana divaricata* cell suspension culture. *J Nat Prod* **54**: 1558–1563.

Davies H. 2009. A role for 'Omics' technologies in food safety sssessment? *Food Control*; doi: 10.1016/j.foodcont.2009.03.002 in press.

Daykin CA, Foxall PJD, Connor SC, Lindon JC, Nicholson JK. 2002. The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by ^{1}H nuclear magnetic resonance spectroscopy. *Anal Biochem* **304**: 220–230.

Defernez M, Colquhoun IJ. 2003. Factors affecting the robustness of metabolite fingerprinting using ^{1}H NMR spectra. *Phytochemistry* **62**: 1009–1017.

Defernez M, Wilson RH. 1997. Infrared spectroscopy: instrumental factors affecting the long-term validity of chemometrics models. *Anal Chem* **69**: 1288–1294.

Dettmer K, Aronov PA, Hammock BD. 2007. Mass spectrometry-based metabolomics. *Mass Spectrom Rev* **26**: 51–78.

Dixon RA, Strack D. 2003. Phytochemistry meets genome analysis, and beyond. *Phytochemistry* **62**: 815–816.

Dixon RA, Gang DR, Charlton AJ, Fiehn O, Kuiper HA, Reynolds TL, Tjeerdema RS, Jeffery EH, German JB, Ridley WP, Seiber JN. 2006. Applications of metabolomics in agriculture. *J Agric Food Chem* **54**: 8984–8994.

Exarchou V, Godejohann M, van Beek TA, Gerothanassis IP, Vervoort J. 2003. LC-UV-solid-phase-extraction-NMR-MS combined with a cryogenic flow probe and its application to the identification of compounds present in Greek oregano. *Anal Chem* **75**: 6288–6294.

Fan TW-M, Lane AN, Shenker M, Bartley JP, Crowley D, Higashi RM. 2001. Comprehensive chemical profiling of gramineous plant root exudates using high-resolution NMR and MS. *Phytochemistry* **57**: 209–221.

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. 2000. Metabolite profiling for plant functional genomics. *Nat Biotechnol* **18**: 1157–1161.

Flores-Sanchez IJ, Pec J, Fei J, Choi YH, Dusek J, Verpoorte R. 2009. Elicitation studies in cell suspension cultures of *Cannabis sativa* L. *J Biotechnol* **143**: 157–168.

Forshed J, Torgrip RJO, Aberg KM, Karlberg B, Lindberg J, Jacobsson SP. 2005. A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *J Pharm Biomed Anal* **38**: 824–832.

Goodacre R. 2005. Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *J Exp Bot* **56**: 245–254.

Goodacre R. 2007. Metabolomics of a superorganism. *J Nutr* **137**: 259S–266S.

Hall RD. 2006. Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol* **169**: 453–468.

Hendrawati O, Yao Q, Kim HK, Linthorst HJM, Erkelens C, Lefeber AWM, Choi YH, Verpoorte R. 2006. Metabolic differentiation of *Arabidopsis* treated with methyl jasmonate using nuclear magnetic resonance spectroscopy. *Plant Sci* **170**: 1118–1124.

Iijima Y, Nakamura Y, Ogata Y, Tanaka K, Sakurai N, Suda K, Suzuki T, Suzuki H, Okazaki K, Kitayama M, Kanaya S, Aoki K, Shibata D. 2008. Metabolite annotations based on the integration of mass spectral information. *Plant J* **54**: 949–962.

Krishnan P, Kruger NJ, Ratcliffe RG. 2005. Metabolite fingerprinting and profiling in plants using NMR. *J Exp Bot* **56**: 255–265.

Leiss KA, Choi YH, Abdel-Farid IB, Verpoorte R, Klinkhamer PGL. 2009. NMR metabolomics of thrips (*Frankliniella occidentalis*) resistance in *Senecio* hybrids. *J Chem Ecol* **35**: 219–229

Liang Y-S, Choi YH, Kim HK, Linthorst HJM, Verpoorte R. 2006a. Metabolomic analysis of methyl jasmonate treated *Brassica rapa* leaves by 2-dimensional NMR spectroscopy. *Phytochemistry* **67**: 2503–2511.

Liang Y-S, Kim HK, Lefeber AWM, Erkelens C, Choi YH, Verpoorte R. 2006b. Identification of phenylpropanoids in *Brassica rapa* leaves treated with methyl jasmonate using two-dimensional nuclear magnetic resonance spectroscopy. *J Chromatogr A* **1112**: 148–155.

Lindon J, Holmes E, Nicholson J. 2001. Pattern recognition methods and applications in biomedical magnetic resonance. *Progr Nuclear Magn Reson Spectr* **39**: 1–40.

Lindon JC, Holmes E, Nicholson JK. 2006. Metabonomics techniques and applications to pharmaceutical research and development. *Pharmaceut Res* **23**: 1075–1088.

Lommen A, Weseman JM, Smith GO, Noteborn HPJM. 1998. On the detection of environmental effects on complex matrices combining off-line liquid chromatography and ^{1}H-NMR. *Biodegradation* **9**: 513–525.

Maher AD, Zirah SFM, Holmes E, Nicholson JK. 2007. Experimental and analytical variation in human urine in ^{1}H NMR spectroscopy-based metabolic phenotyping studies. *Anal Chem* **79**: 5204–5211.

Martin FPJ, Dumas ME, Wang YL, Legido-Quigley C, Yap IKS, Tang HR, Zirah S, Murphy GM, Cloarec O, Lindon JC, Sprenger N, Fay LB, Kochhar S, van Bladeren P, Holmes E, Nicholson JK. 2007. A top-down systems biology view of microbiome-mammalian metabolic interactions in a mouse model. *Mol Syst Biol* **3**: 112.

Mashego MR, Rumbold K, de Mey M, Vandamme E, Soetaert W, Heijnen JJ. 2007. Microbial metabolomics: past, present and future methodologies. *Biotechnol Lett* **29**: 1–16.

Monton MRN, Soga T. 2007. Metabolome analysis by capillary electrophoresis–mass spectrometry. *J Chromatogr A* **1168**: 237–246.

Murch SJ, Rupasinghe HPV, Goodenowe DB, Saxena PK. 2004. A metabolomic analysis of medicinal diversity in Huang-qin (*Scutellaria baicalensis*) genotypes: discovery of novel compounds. *Plant Cell Rep* **23**: 419–425.

Nicholson JK, Lindon JC. 2008. Metabonomics. *Nature* **455**: 1054–1056.

Oliver SG, Winson MK, Kell DB, Baganz F. 1998. Systematic functional analysis of the yeast genome. *Trends Biotechnol* **16**: 373–378.

Pereira GE, Gaudillere J-P, van Leeuwen C, Hilbert G, Lavialle O, Maucourt M, Deborde C, Moing A, Rolin D. 2005. ^{1}H NMR and chemometrics to characterize mature grape berries in four wine-growing areas in Bordeaux, France. *J Agric Food Chem* **53**: 6382–6389.

Rubingh CM, Bijlsma S, Derks EPPA, Bobeldijk I, Verheij ER, Kochhar S, Smilde AK. 2006. Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics* **2**: 53–61.

Sanchez-Sampedro A, Kim HK, Choi YH, Verpoorte R, Corchete P. 2007. Metabolomic alterations in elicitor treated *Silybum marianum* suspension cultures monitored by nuclear magnetic resonance spectroscopy. *J Biotechnol* **130**: 133–142

Sands CJ, Coen M, Maher AD, Ebbels TMD, Holmes E, Lindon JC, Nicholson JK. 2009. Statistical total correlation spectroscopy editing of ^{1}H NMR spectra of biofluids: application to drug metabolite profile identification and enhanced information recovery. *Anal Chem* **81**: 6458–6466

Saude EJ, Slupsky CM, Sykes BD. 2006. Optimization of NMR analysis of biological fluids for quantitative accuracy. *Metabolomics* **2**: 113–123.

Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, van Ommen B, Pujos-Guillot E, Verheij E, Wishart D, Wopereis S. 2009. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, on-line 12 june 2009.

Schauer N, Fernie AR. 2006. Plant metabolomics: towards biological function and mechanism. *Trends Plant Sci* **11**: 508–516.

Schripsema J. 2008. Comprehensive analysis of polar and apolar constituents of butter and margarine by nuclear magnetic resonance, reflecting quality and production processes. *J Agric Food Chem* **56**: 2547–2552.

Schripsema J, Verpoorte R. 1991. Investigation of extracts of plant cell cultures by proton nuclear magnetic resonance spectroscopy. *Phytochem Anal* **2**: 155–162.

Schripsema J, Erkelens C, Verpoorte R. 1991. Intra- and extracellular carbohydrates in plant cell cultures investigated by ^{1}H-NMR. *Plant Cell Rep* **9**: 527–530.

Schripsema J, Caprini GP, van der Heijden R, Bino R, de Vos R, Dagnino D. 2007. Iridoids from *Pentas lanceolata*. *J Nat Prod* **70**: 1495–1498.

Schripsema J, Vianna MD, Rodrigues PAB, de Oliveira JG, Franco RWA. 2009. Metabolomic investigation of fruit flesh gelling of papaya fruit (*Carica papaya* L. 'Golden') by nuclear magnetic resonance and principal component analysis. *Acta Hortic* in press.

Simoh S, Quintana N, Kim HK, Choi YH, Verpoorte R. 2009. Metabolic changes in Agrobacterium tumefaciens-infected *Brassica rapa*. *J Plant Physiol* **166**: 1005–1014

Snyder LR. 1968. *Principles of Adsorption Chromatography*. Marcel Dekker, New York.

Son H-S, Hwang G-S, Kim KM, Ahn H-J, Park W-M, van den Berg F, Hong Y-S, Lee C-H. 2009. Metabolomic studies on geographical grapes and their wines using ^{1}H NMR analysis coupled with multivariate statistics. *J Agric Food Chem* **57**: 1481–1490

Spraul M, Neidig P, Klauck U, Kessler P, Holmes E, Nicholson JK, Sweatman BC, Salman SR, Farrant RD, Rahr E, Beddel CR, Lindon JC. 1994. Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *J Pharm Biomed Anal* **12**: 1215–1225.

Sprogøe K, Stærk D, Ziegler HL, Høgh Jensen T, Holm-Møller SB, Jaroszewski JW. 2008. Combining HPLC-PDA-MS-SPE-NMR with circular dichroism for complete natural product characterization in crude extracts: levorotatory gossypol in *Thespesia danis*. *J Nat Prod* **71**: 516–519.

Tiziani S, Einwas AH, Lodi A, Ludwig C, Bunce CM, Viant MR, Günther UL. 2008. Optimized metabolite extraction from blood serum for ^{1}H nuclear magnetic resonance spectroscopy. *Anal Biochem* **377**: 16–23.

Torgrip RJO, Lindberg J, Linder M, Karlberg B, Jacobsson SP, Kolmert J, Gustafsson I, Schuppe-Koistinen I. 2006. New modes of data partitioning based on PARS peak alignment for improved multivariate biomarker/biopattern detection in ^{1}H-NMR spectroscopic metabolic profiling of urine. *Metabolomics* **2**: 1–19.

Verpoorte R, Choi YH, Kim HK. 2007. NMR-based metabolomics at work in phytochemistry. *Phytochem Rev* **6**: 3–14.

Verpoorte R, Choi YH, Mustafa NR, Kim HK. 2008. Metabolomics: back to basics. *Phytochem Rev* **7**: 525–537.

Viant MR. 2003. Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem Biophys Res Commun* **310**: 943–948.

Vogels J, Tas AC, Venekamp J, van der Greef J. 1996. Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications. *J Chemometr* **10**: 425–438.

Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. 2006. Targeted profiling: Quantitative analysis of ^{1}H NMR metabolomics data. *Anal Chem* **78**: 4430–4442.

Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, van Duijnhoven JPM, van Dorsten FA. 2008. Assessment of PLSDA cross validation. *Metabolomics* **4**: 81–89.

Widarto HT, van der Meijden E, Lefeber AWM, Erkelens C, Kim HK, Choi YH, Verpoorte R. 2006. Metabolomic differentiation of *Brassica rapa* following herbivory by different insect instars using two-dimensional nuclear magnetic resonance spectroscopy. *J Chem Ecol* **32**: 2417–2428

Wold S, Sjostrom M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometr Intell Lab Syst* **58**: 109–130.

Xia J, Bjorndal TC, Tang P, Wishart DS. 2008. MetaboMiner—semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformat*. **9**: 507–522.