J. Pestic. Sci., **32**(3), 200–212 (2007)
DOI: 10.1584/jpestics.R07-03

Review

Use of NMR for metabolic profiling in plant systems

Ian J. COLQUHOUN*

Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, U.K.

(Received November 23, 2006; Accepted February 2, 2007)

The review deals with the applications of solution state ^{1}H NMR to the metabolic profiling of plant tissue extracts. NMR is introduced as one of several measurement techniques that are being used in metabolomics. Samples are measured as extract mixtures without any chromatographic separation of individual compounds. Although a limited quantitative measurement of individual components is feasible, the data analysis generally relies on the application of multivariate statistical methods to the whole spectral traces. Design of experiments and sample preparation and measurement procedures are discussed. Applications are grouped under three major headings: classification and taxonomy; genetically modified plants; chemical treatments, environmental influences and pathogens. The final section introduces some of the newer technologies that will extend the scope of NMR metabolic profiling, including the hyphenation of HPLC with UV, MS and NMR detection. © Pesticide Science Society of Japan

Keywords: NMR, metabolic profiling, plant metabolomics.

Introduction

In principle, proton (^{1}H) NMR can detect any metabolite containing hydrogen. With superconducting magnets operating at frequencies of 400 MHz or above the ^{1}H NMR spectra of biological fluids or tissue extracts are a rich source of qualitative and quantitative information on the compounds present, covering compounds of all chemical classes. NMR has therefore been considered as one of the techniques that can contribute to the emerging field of metabolomics. Metabolomics can be defined as *the quantitative measurement of all low molecular weight metabolites in an organism's cells at a specified time under specific environmental conditions.*

The object of these measurements, the *metabolome*, is made up of many hundreds of metabolites. The plant metabolome is particularly diverse because of the extraordinary variety of secondary metabolites. Estimates for single plants are put at 5,000–10,000 metabolites with a total of maybe 200,000 different structures across the plant kingdom.[1)] Simultaneous measurement of all components of the metabolome by a single high throughput parallel method analogous to those available in transcriptomics and proteomics is complicated by the range of chemical and physical properties of the metabolites and the enormous range of concentrations (pM to mM) at which they occur. Recent reviews are available on the analytical technologies that have been employed in metabolomics[2)] (principally GC/MS and LC/MS in addition to NMR), of applications across the whole field[3)] (including, but not limited to plants) and of NMR in relation to plant metabolomics.[4)]

NMR has both advantages and limitations as a technique for metabolomics. Sample preparation is usually simple and rapid, measurement times are short and readily automated and advanced data analysis methods are available. As well as known compounds (*e.g.* identified from a database) the NMR spectra of complex mixtures may provide sufficient information for the structures of unknown components to be deduced, either from the NMR spectrum of the mixture itself, or after purification. NMR sensitivity is low in comparison with MS so that only compounds to the upper end of the concentration range are detected (typically in the μM to mM range in the NMR tube). The ^{1}H NMR spectra of plant extracts are inevitably crowded because there is a large number of contributing compounds, and most give multiple signals. Each chemically distinct hydrogen atom, or group of hydrogens, has its own chemical shift which gives rise to the dispersion of signals across the spectrum. ^{1}H spectra are also complicated by spin–spin couplings which add to signal multiplicity, although they too are an important source of structural information.

Other important NMR-active nuclei present in biomolecules are ^{13}C, ^{15}N, and ^{31}P. They have greater chemical shift

* To whom correspondence should be addressed.
E-mail: ian.colquhoun@bbsrc.ac.uk
Published online June 20, 2007
© Pesticide Science Society of Japan

ranges than ^{1}H and the major spin–spin couplings (to ^{1}H) are readily removed giving simpler spectra with fewer overlapping signals. However ^{13}C and ^{15}N have low natural abundance and are orders of magnitude less sensitive than ^{1}H, and ^{31}P NMR only detects a relatively small number of compounds. ^{1}H is therefore the preferred nucleus for NMR studies of plant extracts and high magnetic fields give the best sensitivity and signal dispersion.

Analysis of biological samples as crude extracts without any separation step is known as 'fingerprinting' or 'profiling.' Other spectroscopic fingerprinting approaches[2] include direct injection mass spectrometry (DIMS) and Fourier-transform infrared (FTIR) spectroscopy. Plant extracts are very complex in composition and, if many samples are examined, it is difficult to make meaningful comparisons of large numbers of spectra or chromatograms 'by eye.' Multivariate statistical methods can be extremely useful, as they are able to compress data into a more easily managed form. This can assist in visualizing how a given sample relates to other samples—for example experimental samples relative to controls.

In this article we review applications of the NMR fingerprinting—multivariate analysis approach to plant systems. We indicate where signal assignments are available for plant species and cover applications of NMR profiling to classification, transgenic plants, effects of environment (including herbicide treatments and pathogen interactions), and discuss some of the recently developed 'hyphenated' technologies. We do not attempt to cover the interesting but rather independent areas of '*in vivo*' NMR or flux measurements.[5]

Experimental Design, Sample Preparation and Signal Assignment

Metabolomics and metabolite profiling experiments involve quantitative comparisons of the levels of multiple analytes across two or more groups of samples. Some studies use these measurements to classify the samples as members of previously known (or suspected) groups. Other studies aim to detect the effects of a particular treatment applied to one group of samples but not to an otherwise identical control group. In the first case, typically applied to samples such as plant based foods or herbal medicines, the experimental design would aim to cover variations across the population (different varieties, growing seasons, regions of origin and so on). This requires some expert knowledge of the product concerned but may be subject to practical constraints in terms of the number of samples it is feasible to collect.

Secondly, plants can be specifically grown for a metabolomics study and the experimental design is then more directly under control of the investigator. However to establish whether there are significant differences between the treated and control plants an adequate number of replicates is essential. Replicates may be individual plants (common in GC/MS metabolomics experiments) or pooled material from several plants (more usual in NMR investigations). Even plants that are grown under nominally 'identical' conditions in a controlled environment chamber show considerable plant to plant variation in metabolite composition. The biological variability is increased further when plants grown at different times are to be compared.

Lewis *et al.*[6] carried out a ^{1}H NMR investigation of *Arabidopsis* extracts from plants grown in a controlled environment. They found that extracts from individual plants showed approximately twice as much variation as pooled extracts combining material from all 24 plants in a tray. They established a protocol where one biological replicate consists of the combined freeze dried material from a tray of 24 plants and measure three such replicates for each treatment/ line (see http://www.metabolomics.bbsrc.ac.uk/techniques.htm for further details). Standardisation should be applied wherever possible to factors such as the development stage and time of day at which plants are harvested. Other sources of variation that can be introduced are from the sample preparation procedure and the analytical measurements. These need to be minimised but are generally less important than the natural biological variation.

Different sample preparation procedures have been employed depending on the type of sample. Leaf material is generally immersed in liquid nitrogen immediately after harvesting and stored frozen (−80°C) or else freeze dried and stored until required for extraction.[6,7] Potato tubers may be stored (10°C) in the dark for two weeks after harvest and then a special sampling procedure is recommended to minimise variations from metabolic concentration gradients within the tuber.[8] Fruit samples may also be examined as juices by standard high resolution NMR or as pulps (by ^{1}H high resolution spinning NMR, HRMAS[9]).

Various solvents and solvent combinations have been used for extraction of plant tissues, depending on whether the main interest lies in the polar or non-polar constituents or in both. The NMR technique itself requires the presence of some deuterated solvent to provide a field-frequency lock signal and a signal on which resolution adjustment can be carried out for each sample. It is advisable to buffer aqueous or part-aqueous solvents because the chemical shifts of many compounds are sensitive to pH and if the resulting inter-sample differences are not minimised it will lead to difficulties when multivariate analyses are attempted. Phosphate buffers are most commonly used as they do not give any additional ^{1}H signals in the spectrum. Otherwise the pH of all samples can be adjusted to a common value by adding small volumes of hydrochloric acid or sodium hydroxide to the solutions but this is necessarily a slower procedure.[10]

A weighed amount of sample (often 15–30 mg for freeze dried powders) is extracted into a deuterated solvent (~1 mL) with stirring, vortexing or sonication. After centrifugation, a measured amount of supernatant (400–750 μL) is transferred to an NMR tube and spectra may be obtained immediately, making this one of the simplest of all metabolomics prepara-

tion procedures. Solvents that have been used include 400 mM D_2O phosphate buffer[11]; 70% d_4-methanol/ 30% D_2O (with 100 mM phosphate buffer)[12]; 20% d_4-methanol/ 80% D_2O[13] for polar extractions. Lewis *et al.*[6] used two heating steps (50°/10 min and 90°/2 min), the second step being included to ensure enzyme inactivation but this does not appear to be necessary with 70% methanol and the extraction may be done at room temperature. Chloroform/methanol/water mixtures in various proportions followed by phase separation of organic and aqueous layers have most often been used for comprehensive extraction of polar and non-polar compounds.[14] After solvent removal the organic residue is dissolved in *d*-chloroform for NMR of the non-polar fraction. Moing *et al.* have employed more elaborate procedures involving repeated extraction and freeze drying steps and removal of metal ions because their aim was to develop an NMR method for the absolute quantification of polar metabolites.[15] A procedure using a solid phase extraction step and elution solvents of increasing hydrophobicity was developed for fractionation of the aqueous extract of tomato.[16] It increased comprehensiveness by enabling NMR detection of some components in the fractions that would not have been measurable in the total extract, but the quantitative procedure is rather lengthy for routine use.

NMR spectra are measured, usually with the simple presaturation pulse sequence (low power irradiation at the water signal frequency during the relaxation delay) for samples in D_2O or the NOESY-presaturation sequence for samples such as juices with a high proportion of H_2O. Acquisition times (2–3 sec) and relaxation delays (2–3 sec) are commonly set to give a recycle delay of ~5 sec and total acquisition times of ~15 min. If the aim is absolute quantification of all metabolites within a spectrum a longer relaxation delay (~20 sec) should be used. The shorter relaxation delays give spectra that are suitable for comparison of the amounts of the same metabolite in different spectra. The longer delay should be used to obtain accurate relative intensities of different metabolites within the same spectrum.[15] Provided that the measurement conditions are unchanged the NMR spectra obtained with modern instruments are highly repeatable.

Samples are generally run under automation in batches of up to 60 tubes at a time (more samples can be handled in a batch by spectrometers equipped with flow injection probes). The probe is tuned on the first sample of a batch: it can be tuned on every sample within automation if the probe is fitted with a suitable accessory. Tuning on the first sample will generally be sufficient when all samples have similar ionic strengths. It is convenient to keep the receiver gain at the same level for all samples but if this is not possible a correction can be made subsequently. Samples should be run at a fixed controlled temperature, not at the ambient temperature, and should be allowed time to equilibrate after being loaded into the spectrometer. The line-width and line-shape (resolution) should be identical for all samples. This condition is usually well met on modern instruments with gradient assisted shimming carried out on each sample.

Many signals can be assigned by comparison with libraries of reference compounds, or by two-dimensional NMR. Substantial tables of assignments have been published for a number of samples of plant origin including apple,[17] mango,[9] orange,[10] tomato (fruit[12,18] and root[15]), potato,[8] strawberry,[15] *Arabidopsis* (leaf[7]) and lettuce (leaf[19]). Most of these assignments were for aqueous samples or polar extracts and many compounds are common to all samples (*e.g.* amino acids, organic acids, simple sugars). Major secondary metabolites, for example phenylpropanoids and glucosinolates in *Arabidopsis*, are readily detected.[7]

The study on lettuce[19] was particularly comprehensive in that both polar and non-polar extracts were assigned and a comprehensive set of 2D NMR experiments (including diffusion ordered spectroscopy, DOSY) was carried out. Metabolites identified in the aqueous extract included inulins ($GlcFru_n$ where $n=2$–5, the degree of polymerisation being estimated by DOSY), mono and di-caffeoyl substituted organic acids including dicaffeoyltartaric acid (chicoric acid) and chlorogenic acid. Compounds identified in the non-polar extract included pheophytins (chlorophyll structure), carotenoids (lutein, β-carotene) and sterols (β-sitosterol and stigmasterol).

The assignment of remaining compounds was simplified by isolating an 'acetone insoluble' fraction of the non-polar extract in which galactosylglycerols, sulpholipids, and phospholipids were detected and the percentages of free fatty acids and di- and poly-unsaturated fatty acid chains were calculated.

Multivariate Data Analysis

Chemometrics is the area of mathematics and computing in which data processing tools and multivariate statistical techniques are applied to the high dimensional data produced by modern analytical techniques.[20,21] An NMR spectrum consists of intensities at thousands of data points across the chemical shift scale. The chemical shift values at which these intensities are measured are called the variables (or variates).

The variables can be data points taken directly from the spectrum[10] or the number of variables may be reduced by a so-called bucketing procedure. This involves dividing the spectrum up into bins of a certain width and summing all the intensities within each bin.[22] The result is that the number of variables is reduced from say 16,000 to about 250 if a typical bin width of 0.04 ppm is used. Although bucketing has been widely adopted (it has advantages and disadvantages) it is not a computational requirement since desktop PCs can now cope easily with the full data sets. It is also possible (but not so straightforward) to work with variables that are related directly to concentrations of individual compounds[15] by using integrated peak intensities or line shape fits.

Spectra are then assembled in a table (data matrix) in which

the rows correspond to the samples (observations) and the columns to the chemical shift values (variables). Values in each column are usually 'mean-centred' (the mean of the column is calculated and subtracted from each value). Columns may also be 'autoscaled' (the mean is subtracted and the values are divided by the standard deviation for that column; each column then has a mean of zero and variance of unity). The covariance matrix is obtained by multiplying the transpose of the mean-centred matrix by the mean-centred matrix itself. In the case of autoscaling the same operation gives the correlation matrix.[20)]

Metabolomics experiments often have the aim of classifying samples into different groups (*e.g.* genotypes or treatments). Both univariate and multivariate statistical methods have a role in distinguishing between groups and in determining where in the profiles the differences lie. Univariate statistical tests (*t*-test, ANOVA) can be carried out on one variable (column) at a time but many tests are required. With multivariate methods all variables are considered simultaneously. The multivariate data for each sample (row) constitutes a vector. Samples can be pictured conceptually as points in multidimensional space with each variable as an axis; the location of the sample is determined by the intensities which are the co-ordinates on each axis. This leads to the idea of using the distances between samples as a basis for their classification into groups.[20)]

Principal component analysis (PCA) is a multivariate method that can drastically reduce the number of variables needed to describe the variance in the data set.[20,21)] Mathematically, PCA involves determining the eigenvectors of the covariance (or correlation) matrix. In graphical terms PCA generates a rotated set of axes using linear combinations of the original axes. The new axes are calculated so that the first principal component, PC1, defines the direction of maximum spread (variance) in the data. The second (orthogonal) axis, PC2, defines the direction of greatest remaining spread and so on. The **scores** are the co-ordinates of the samples in the new axis system defined by the PCs (mathematically the scores are obtained by multiplying the data matrix by the eigenvectors). The most important PCs are the first ones to be calculated in the sense that they account for the greatest amount of variance (typically the first two or three PCs may account for over half the total variance in the data set). This property means that we can replace the hundreds or thousands of original variables with a new set of variables (typically <20), the principal components, without loss of information. When groups of samples have systematic differences in the concentrations of major compounds then **scores plots** (*e.g.* PC1 *vs.* PC2, PC3 *vs.* PC4…) on one or more principal components show spatial clustering.

The PC **loadings** (another term for the eigenvectors) show the contribution of the original variables to each PC. **Loadings plots** identify the data points with high loadings and, since the loadings closely resemble the original NMR spectra (at least with the covariance method) the compounds that are responsible for the differences between groups can be identified. The decomposition of the data matrix into scores and loadings matrices and the appearance of the associated plots are shown schematically in Fig. 1.

The correlation matrix method is useful when the classification relies more on differences between compounds present in low concentrations than on the major compounds. Autoscaling increases the influence of weaker signals but the loadings are not readily interpretable; Pareto scaling[23)] (division of each column by the square root of the standard deviation) is an intermediate pre-processing method that has been much used with NMR data because it gives some emphasis to weaker signals but still provides interpretable loadings.

PCA is an exploratory or unsupervised method, so-called because the experimental data alone is analysed. Supervised classification methods may be appropriate when biological variation is more prominent than systematic differences between groups. In that case the group separation may not be easily visualised from two-dimensional scores plots of the first few pairs of PCs. In a supervised method the experimental data and group membership are supplied together as separate tables and the group information influences calculation of the scores. One of the most popular supervised methods is **Partial Least Squares—Discriminant Analysis (PLS-DA)** which, like PCA, relies on a rotation of axes.[20,23)] However the axis directions are calculated so that the resulting sample scores give the optimal discrimination between groups rather than being based solely on the spread in the experimental data. Scores and loading plots from a PLS-DA carried out on NMR spectra of transgenic and control tomato fruits at three

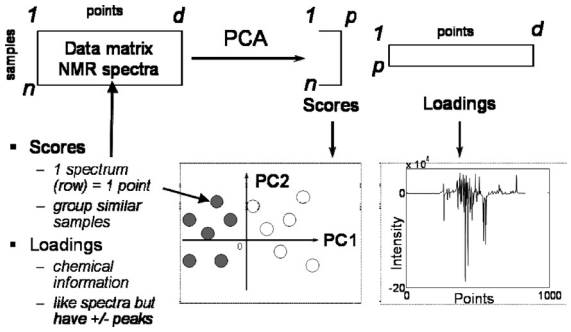

Fig. 1. Schematic depiction of data compression by PCA. An $n \times d$ data matrix (n spectra, each with d data points or bins) is replaced by an $n \times p$ scores matrix where $p \ll d$ (p is the number of principal components retained). The original spectra and the scores are related to each other by the loadings which give the contribution of each of the d data points to each PC. Values in one column of the scores matrix may be plotted *versus* values in another column (*e.g.* PC1 *vs.* PC2) to visualise any sample clustering. The group of samples with negative scores on PC1 (filled circles) will have higher average levels of compounds with negative signals on the PC1 loadings plot shown whilst the group with positive scores (open circles) will have higher levels of compounds giving positive loading signals.

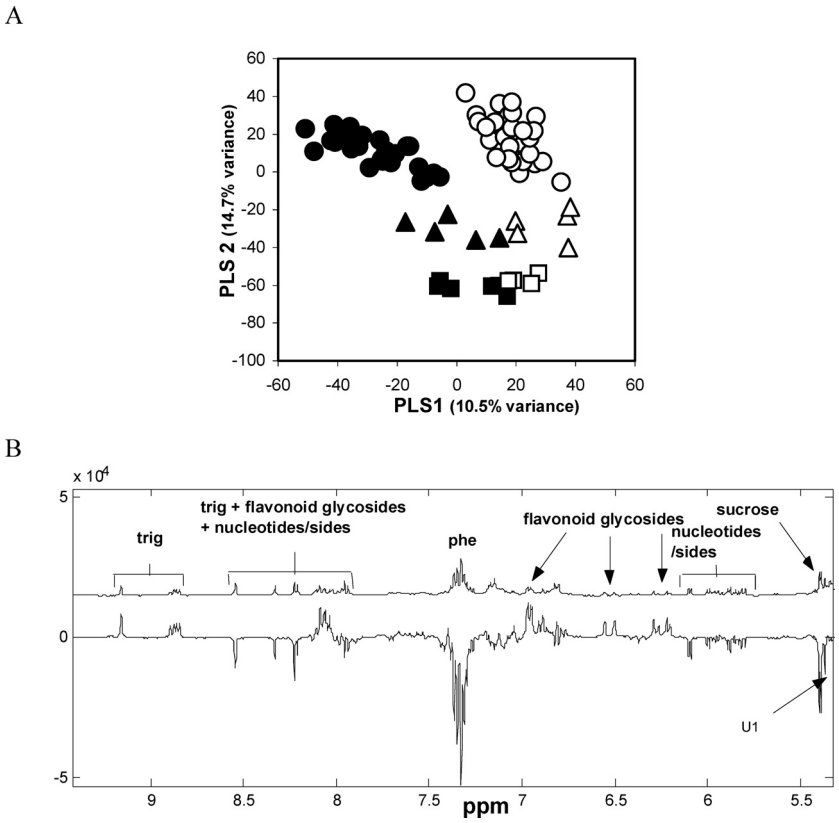

Fig. 2. (A) PLS scores (first two axes) of NMR data set from extracts of high flavonoid transgenic tomatoes and azygous controls. Open and filled symbols represent transgenic and control tomatoes respectively. Squares, triangles and circles represent green, turning and red ripe tomatoes. (B) PLS loading 1 (lower trace) and mean spectrum for red ripe transgenics (upper trace). For simplicity only the low-field part of the spectrum is displayed although calculations were based on the full spectrum. Key: trig, trigonelline; phe, phenylalanine. Together the scores and loadings plots indicate (for example) increased levels of flavonoid glycosides and trig but a decreased level of phe in the transgenics. Adapted with permission from *J. Agric. Food Chem.* **51**, 2447. Copyright (2003) American Chemical Society.

different stages of ripening are shown in Fig. 2.

Many of the applications of multivariate analysis described in this review have stopped with an examination of two-dimensional scores plots and the associated loadings. However it is also possible to establish a PC or PLS model by deciding how many principal components or PLS factors to retain. The number of factors is chosen by analysing a set of samples of known origin and determining how many factors are needed to get a high (preferably 100%) classification success rate. Samples may be classified with a distance-based method such as **Linear Discriminant Analysis (LDA)** with the PC scores as the inputs.[20] **SIMCA** is a PCA based classification method in which only one group of samples is modelled at a time and statistical tests are then applied to determine whether further samples should be included in the group or not.[20] If there are a number of groups present a different PCA model is constructed for each group.

An application of **Artificial Neural Networks (ANNs)** to classification of NMR profiles is described below. With supervised methods (PLS-DA, SIMCA, ANN) it is important to check that the data are not 'over-fit': overfitting is most likely to occur when experiments are under-determined *i.e.* when there is an extremely large number of variates (as in the case of NMR) and a small number of samples. An 'over-fit' model is one that appears to classify correctly all the samples on which it is based but which does no better than random guesses when presented with new samples to predict. The answer to this problem is to use only a proportion of the total samples available to build the model and then to test and validate it using the remaining samples.[20] The procedure is described for the example of ANNs in the section on 'Chemical treatments, environmental influences and pathogens,' although it can be applied equally to other supervised methods. Further confidence is gained when interpretation of the loadings plots leads to a plausible biochemical explanation of the differences between samples.

Classification and Taxonomy

NMR spectroscopy in combination with multivariate analysis has been applied to a number of classification problems including species and cultivar discrimination and quality assessment of foods and herbal medicines (food authenticity appli-

cations have been reviewed[24]).

Leaves of *Ilex paraguariensis* (Yerba mate) are the source of mate, a tea like drink consumed in S. America. NMR profiling has been carried out to discriminate between *I. paraguariensis* and 10 other *Ilex* species.[14] Four samples were analysed for each species, grown from seeds collected in different years and locations. A methanol/water/chloroform extraction was carried out on dried leaves and both organic and aqueous fractions were examined. The organic fraction of *I. paraguariensis* contained caffeine and theobromine but these were not present in any of the other *Ilex* species. Signals of fatty acids and a triterpene, probably ursolic acid, were identified in all samples. PCA of the organic fraction NMR data readily separated *I. paraguariensis* samples from other *Ilex* species by virtue of the caffeine content, and partly separated other species from each other after removal of *I. paraguariensis* samples. Prominent secondary metabolites in the aqueous fraction were phenylpropanoids and, in some species, arbutin (hydroquinone-β-glucoside) which had not previously been associated with *Ilex* species. PCA of aqueous fraction data separated *I. paraguariensis* and three other species using the first 3 PC axes and the loadings plots showed that differences in levels of phenylpropanoids, arbutin and sucrose were mainly responsible for the separation. SIMCA models based on 4–7 PCs were constructed for each species from separate treatments of organic and aqueous fraction data. A successful model (no confusion between the modelled species and other samples) could be established for every species in the case of the aqueous fraction but three of the species were confused when organic fraction data was used. A similar approach was used to discriminate between *Strychnos* species,[25] small trees that are a source of strychnine and other alkaloids. Methanol extracts (three species, but with different plant tissues giving 8 groups in all) were examined by ^{1}H NMR. In addition to carbohydrates and fatty acids the compounds identified were alkaloids (including strychnine and brucine), caffeic acid esters and (in seeds) loganin. Signals of alkaloids (not strychnine), fatty acids and loganin were important for discrimination between the groups.

In another study of this type NMR fingerprinting was applied to discrimination of three *Ephedra* species,[26] including *Ephedra sinica*, the main source of ephedrine and pseudoephedrine alkaloids. Nine commercial samples of *Ephedra* medicinal herbs were also included. The organic fraction was dominated by ephedrine and pseudoephedrine signals. The three species were easily separated by PCA thanks to their different mean levels of these compounds but little other chemical information was available from this fraction. The aromatic region of the NMR spectrum was quite different for the three species in the aqueous extracts. Compounds identified included a 'benzoic acid analogue,' phenylalanine and phenypropanoids. All three species were separated by PCA (on PC1). One commercial sample appeared close to *E. intermedia* whilst another was apparently a mixture of *E. intermedia* and *E. sinica*. The remaining commercial samples were separated from all three authenticated species on PC1 and according to the PC1 loading had higher levels of the benzoic acid analogue than any of the individual species.

In a novel application NMR profiling was carried out on two *Senecio* (ragwort) species and their F_1 hybrids[27] (aqueous methanolic extracts from aerial tissue of 8 individual plants each of *S. aquaticus* (SA), *S. jacobaea* (SJ) and the F_1 hybrids (H)). The aim was to determine whether levels of existing metabolites were significantly altered in the hybrids and even whether any completely new metabolites were introduced. PCA was not carried out on the standard one-dimensional NMR spectra but on the projections from two-dimensional J-resolved spectra which have a recording time of about 15 min per spectrum. The projection on the chemical shift axis is simplified in comparison with a conventional ^{1}H spectrum in that the J-coupling is removed, and this in turn should simplify interpretation of the PC loadings plots.

The three groups of samples showed clustering in a scores plot (PC1 *vs.* PC3) with SJ separated from SA and H on PC1 and H from SA on PC3. Examination of corresponding loadings suggested a variety of common and less familiar metabolites were responsible for the separation including Ala, chlorogenic acid, flavonoids, fumaric, malic and succinic acids, Glc, Suc, jacaranone analogues and pyrrolizidine alkaloids (PA). Statistical analysis of individual signals for the above compounds showed that mean levels of metabolites in the hybrids were either intermediate between species or lower than both. Several compounds picked out from the loadings did not show significant differences when examined by ANOVA (based on SA, SJ and hybrid groups). No novel metabolites were revealed in the hybrids but this might be because of the sensitivity limitations of the technique or because only partial structural information is available (*e.g.* PAs were identified by a characteristic signal from the core structure but individual variants, known to be present were not distinguished).

NMR profiling showed some ability to distinguish green teas (70% methanol extracts) by country of origin but the sample collection was biased towards Chinese teas with relatively few examples from each of the other countries.[28] More convincing was the discrimination of Chinese Longjing teas (high quality) from teas of lower grade. Principal component loadings indicated that compounds including theanine, theogallin, gallic acid, caffeine and various catechins were present in higher amounts in the Longjing teas. ANOVA was used to confirm the conclusions for individual compounds. The feasibility of using NMR for quality control of phytomedicines has been explored. An examination of 14 commercial feverfew preparations[29] (aqueous and organic extracts from the commercial tablets) showed that two of them had a very different composition from the other 12. The remaining 12 samples could mostly be distinguished from one another although the differences were smaller and some

groups of 2 or 3 samples were apparent. Different batch numbers of a product from the same supplier could be distinguished. NMR and PCA could distinguish chamomile flower samples[30] from three countries as well as readily distinguishing between extracts from flowers alone and mixtures with added stalk. Different extraction methods were tested including a boiling water infusion, a high pressure extraction with water or aqueous ethanol and NMR samples were prepared with and without an intermediate drying step. Compounds identified included amino acids, chlorogenic acid and α-bisabolol. PCA also showed differences in extraction efficiency between water (which favoured sugars) and aqueous ethanol (favoured amino and organic acids). Drying the aqueous extract reduced the intensity of several signals including acetate. Apart from this case, systematic comparisons of sample preparation protocols for NMR profiling have not been widely reported.

Genetically Modified Plants

Metabolomics methods, including NMR, have been suggested as a means of extending current substantial equivalence procedures for the safety testing of transgenic food crops.[31] The purpose of substantial equivalence testing is to check that no unintended changes in composition have occurred as a result of the genetic modification.[32] It is based on the statistical comparison of the concentrations of a pre-selected set of compounds in the GM line and in appropriate non-GM controls that are regarded as safe. The selection of compounds includes important nutrients and known natural toxicants for the crop in question and is intended to be a sufficient basis for a pragmatic conclusion to be drawn about the equivalence or non-equivalence of GM and conventional crops. It has been criticised on the grounds that it can never detect unexpected changes that fall outside the core group of analysed metabolites.[33] Since metabolomics does not involve any pre-selection of compounds for analysis it has been proposed that its use could add to the comprehensiveness of testing and enable detection of unexpected changes, should they occur. Several publications have explored the possibility of using NMR profiling in this context.

In an early application of the method NMR was used to determine differences between GM lines and controls in two series of modified tomatoes.[16] Both aqueous and organic extracts were prepared from relatively large amounts (50 g and 20 g fresh weight, respectively) of tomato fruit. In an attempt to increase the range of compounds detected the aqueous extract was separated into four fractions containing compounds of different hydrophobicity by solid phase extraction. This separated compounds such as the abundant sugars, amino and organic acids from less concentrated metabolites such as glycoalkaloids. Using customised peak picking and data alignment software it was possible to list some 2500–3000 signals (over the five fractions) for each sample. Note that the number of compounds is likely to be less than one-tenth of this total although no attempt was made to identify the compounds. Signal amplitudes were measured and systematic statistical testing was carried out to test for significant differences between various groups of samples.

One of the modifications (in which the Cry1Ab5 protein from *Bacillus thuringiensis* was expressed) was not expected to cause changes in the tomato fruit composition. A relatively large number of significant differences (100–200 signals with $p<0.01$) was detected between these Bt-tomatoes and controls in each of the three years that the crops were grown, but no significant differences at all were obtained when the data from all three years were combined. This means that the differences detected within each year's crop, mostly no more than two-fold changes of the mean, were not consistently present and could be a reflection of environmental perturbations or simply a statistical consequence of making a very large number of comparisons.[16]

A delayed ripening GM line, which might have been expected to show greater changes than the Bt-tomatoes was also examined.[16] It too showed a marked diminution in the number of signals showing significant differences (from 249 to 26) as the number of control batches was increased from one to three. It would appear unwise to base decisions about equivalence purely on 'fingerprinting' and statistics without considering the origin of the significant differences and their possible biological origin and relevance.

In many cases it is necessary to use NMR in conjunction with other techniques. Genetic modification of tomato by simultaneous overexpression of two transcription factors from maize greatly enhanced the content of flavonoid glycosides in the tomato fruit flesh. ^{1}H NMR profiling was carried out on the GM tomatoes and matched azygous controls (grown hydroponically in a greenhouse under the same conditions) in order to detect compositional changes affecting both flavonoids and other types of compound.[12] GM and control fruits were analysed at different ripening stages (green, turning, red ripe) by ^{1}H NMR of 70% methanol extracts. PCA and PLS scores plots (Fig. 2A) from the resulting spectra showed separation of samples into groups according to whether they were GM or control (seen on PLS factor 1) and according to ripening stage (PLS factor 2). The separation of GM and control groups increased through the three ripening stages. Spectra of representative samples were assigned as far as possible and examination of the PLS loadings (Fig. 2B) together with comparison of mean spectra for the different groups was then used to determine which compounds were responsible for the discrimination seen in the scores plot.

The increased level of flavonoid glycosides in the GM tomatoes was largely responsible for the discrimination between transgenics and controls. Many of the flavonoids were detected as novel signals in the GM samples. ANOVA on selected peaks of about 20 non-flavonoid compounds identified from the loadings (*e.g.* sucrose, phenylalanine and trigonelline) also showed statistically significant differences

between GM and control groups in the red tomatoes. In contrast to the flavonoids however none of these compounds showed changes in mean levels that were greater than two- or three-fold. Such changes were considered small in relation to the background of natural variability and confirmed that the effects of the modification were essentially confined to the targeted pathway.[12)]

The flavonoid signals themselves could only be partly assigned from the ^{1}H NMR spectra of the unfractionated extracts. A fuller characterisation was achieved using a combination of LC/NMR, LC/MS and LC/UV experiments[34)] from which nine major flavonoids were identified in the GM tomatoes, mostly kaempferol glycosides that were known previously in other plants but not in tomato. Only three of the nine compounds could be detected in control tomatoes.

Metabolite profiling has also been used to assess compositional changes occurring in potato tubers after genetic modifications had been made to different metabolic pathways.[8)] The techniques used were ^{1}H NMR and LC/UV. Results were largely complementary since only a few compounds (aromatic amino acids, chlorogenic acid) were detected by both methods. The samples came from 4 groups with modifications to primary carbon metabolism (cv. Record), starch synthesis, glycoprotein processing or polyamine/ethylene metabolism (all cv. Desirée). Each group was represented by several independent lines with wild-type, empty vector or tissue culture controls, all grown together under containment (polytunnel). Differences in composition were sought at the level of whole profiles (by PCA) or individual compounds (by ANOVA).

Unlike the high-flavonoid tomatoes no single compound or class of compounds was dramatically enhanced (or reduced) by these modifications. The most significant differences between GM and control tubers were found in two of the four Desirée lines with modified polyamine metabolism. Compounds affected included proline, trigonelline and several phenolics. However that modification produced an abnormal phenotype and the changes observed were most likely associated with a stress response. Certain lines from the other groups had several compounds present in significantly higher or lower amounts than the controls, but the differences in mean values never amounted to more than a 2–3 fold change. Against the background of variability in the whole data set such changes were not deemed important: the differences between the two cvs. Record and Desirée were greater than differences found between GM and controls for either cultivar.[8)]

NMR profiling with multivariate analysis has also been used in an attempt to detect unintended effects of GM in transgenic peas[35)] (samples were aqueous extracts from combined leaf material of individual plants). Six independent T-DNA insertion lines (with one to four insertions per line) were compared with each other and with wild type and null segregant controls. In an initial experiment PCA-linear discriminant analysis successfully separated one transgenic line (T_3 generation) from WT controls. With a larger data set (T_4 generation) on plants from all six lines and two controls PCA-LDA was not successful in separating transgenics and controls. Also an ANOVA type procedure based on the 86 highest variance standardised points in the NMR spectra found no significant differences between the group of (all) T_4 transgenics and the group of (all) controls. It was possible to discriminate any individual insertion line from either type of control by PLS-LDA when pair-wise comparisons were made, but the differences cannot be consistent across the lines. The authors argued that the main difference between the metabolic profiles of the transgenics and the WT was a reduced variability in the transgenics, but that this variability was gradually restored as successive generations became removed from the original transformation.

Two bacterial genes encoding enzymes involved in the salicylic acid biosynthetic pathway were introduced into tobacco to give constitutive salicylic acid producing (CSA) plants. NMR profiles of the leaves and veins of CSA and WT plants have been obtained.[36)] Both types of plant were inoculated with tobacco mosaic virus as salicylic acid is a signalling molecule involved in the systemic acquired resistance of plants to TMV. After 10 days samples were prepared from leaves and veins using both the inoculated leaves and the (systemic) leaves directly above the inoculated ones. PCA of NMR spectra from all samples showed four groups separated by their PC1 and PC2 scores. The separation on PC1 was between leaf and vein tissues and that on PC2 between WT and CSA samples. The WT samples were more dispersed than CSA regarding both the differences between non-inoculated, inoculated and systemic leaves and the variation between plant replicates. This is probably because the CSA plants have increased resistance to the viral infection and their profiles are less perturbed. The loadings identified decreased levels of Suc, Glc and chlorogenic acid and increased malic acid and alanine in CSA as compared with WT plants. In the WT chlorogenic acid was reduced in inoculated leaves compared with non-inoculated and systemic leaves. Possible reasons for these differences were proposed.

Other examples of the application of NMR profiling and multivariate analysis to various tissues of transgenic plants include Bt-maize (seeds),[37)] *Arabidopsis* transformed with an antisense chalcone synthase gene[7)] (leaves), tomato overexpressing an *Arabidopsis* hexokinase gene[15)] (roots) and *Arabidopsis* transformed with a *Sorghum* phospho*enol*pyruvate carboxylase gene but showing decreased PEPC activity[15)] (leaves). One of the great promises of metabolomics is that it will come to play an important role in functional genomics: the two examples mentioned above[15)] represent the first attempt to use NMR profiling of plants for this purpose.

Chemical Treatments, Environmental Influences and Pathogens

A number of studies have used NMR profiling to investigate

effects of various environmental treatments on metabolite composition including application of herbicides; effects of metal ions; treatment with signalling compounds associated with stress response; and infection with pathogens.

In two proof of concept papers it was shown that NMR profiles of plant extracts from maize seedlings treated with different herbicides could be distinguished according to the mode of action (MOA) of the herbicide.[38,39] This holds promise for screening of novel compounds if the MOA of the novel compound is the same as that of a known compound already investigated. The potential utility of such an approach is directly related to the comprehensiveness of the reference data base available.

In a much more extensive follow-up study by one of the groups an attempt was made to build such a data-base.[40] Maize seedlings were treated with one of 27 herbicides representing 19 known MOAs (some MOAs were represented by several compounds). In this large-scale study emphasis was placed on having a large number of biological replicates, on achieving good repeatability through growth of the seedlings in a liquid medium in a controlled environment chamber, and on developing a highly controlled sample preparation and NMR measurement procedure. Altogether over 400 samples, including controls, were analysed as aqueous acidic extracts of meristematic tissue.

Between 10 and 20 artificial neural networks were constructed for classification of samples according to MOA with additional categories for *controls* and *unknowns*, the last category allowing for the possibility that the MOA of a novel compound might not correspond to any of those used to create the network. Input to the network consisted of the NMR intensities from individual samples reduced to 1080 data points (from 16 K in the original spectra) with total intensity of each spectrum normalised to unity. Each network was constructed from a different random allocation of samples between three roughly equivalent sized sets, the training, validation and test sets. Test set samples were never actually used for construction of the network, only to test the quality of the predictions. Classification success rates for the test samples were then reported as averages for all the networks constructed. The result could be 'correct,' 'incorrect' or 'unknown.'

The first overall model to be examined used all the data available covering all MOAs in the design. In the sense used here 'model' refers to which groups or MOAs are included/excluded for the calculation of the ANNs. Overall results for the all-inclusive model were 64% correct, 6% incorrect, 30% unknown. About 5 groups were extremely well predicted with over 90% correct. The well identified groups were removed from a second round of modelling which proceeded as before but included only those groups with more than one-third of samples classified as unknown or mis-assigned. Limiting the model in this way improved the classification success rate for the 'difficult' groups compared with the all-inclusive model but of equal interest was the summarising of the results in the form of a confusion matrix which splits up the incorrect assignments to show exactly how the mis-assignments are allocated. This is an economical way of summarising the findings when many groups are involved and can suggest which pathways are most closely related (at least in terms of eventual metabolite profiles) and which are unrelated.

The authors also subjected their models to various 'leave one out' tests that aimed to reproduce situations likely to arise in the screening of real unknown compounds. For example all replicates for a particular compound were excluded when training either the inclusive or the more limited model. This included the case where the excluded compound was one of several with the same MOA as well as the case where that compound was the only member of its class. In the first case one would expect the assignment to be correct but in the second case 'unknown.' The authors discuss the extent to which these expectations are met and suggest possible reasons why the assignments are not always as expected. There are many lessons to be learned from this paper[40] on the organisation and data analysis of complex metabolic profiling studies with multiple classification groups and large numbers of samples.

In an attempt to apply the same method a phytotoxin (pyrenophorol) isolated from a pathogen with host specificity for *Avena sterilis* (wild oat) was applied to seedlings of *A. sterilis* and the NMR profile compared with those from controls and seedlings treated with six herbicides with different MOAs.[41] However multivariate analysis (by PCA, PLS-DA and SIMCA) showed that although pyrenophorol treated samples were clearly differentiated from controls, they did not co-classify with any of the selected herbicide treatments.

NMR and GC/MS profiling has allowed components of barley root exudates to be characterised and measured, including several derivatives of mugineic acid (phytosiderophores or specialised iron-complexing ligands). Plants were grown in nutrient solutions subject to increasing levels of iron deficiency which resulted in increased production of exudate compounds with a high proportion of phytosiderophore.[42] Further studies of barley and wheat exudates showed that iron deficiency increased production of malate and mugineic acids and resulted in increased uptake of some metals (Cu, Mn, Zn), but not cadmium.[43] The species *Silenus cucubalus* is tolerant to uptake of Cd^{2+} which is chelated by a family of peptide ligands in the plant. Overall metabolic effects of Cd treatment were studied[44] by NMR profiling of *S. cucubalus* cell extracts from cell cultures grown in media with and without added Cd. Uptake of Cd decreased levels of glutamine and branched chain amino acids but increased, among others, malic acid, acetate and glucose.

Treatment of *Arabidopsis* and *Brassica rapa* (turnip) with methyl jasmonate has been used to stimulate plant defence responses and the resulting NMR profiles have been examined. In *B. rapa* the emphasis was on the identification and NMR assignment of phenylpropanoids such as caffeoyl-,

coumaroyl-, sinapoyl-malate *etc.* which are produced in response to the jasmonate treatment.[45] In *Arabidopsis* (Col-0 ecotype) major phenolic metabolites, mainly kaempferol and quercetin glycosides and sinapoyl-malate, were identified after concentration and fractionation.[46] Changes in metabolite composition with time (0–168 hr) after methyl jasmonate treatment were then followed by NMR, using two-dimensional J-resolved spectra of extracts in aqueous methanol. Projection of the 2D spectrum onto the chemical shift axis gives a pseudo 'proton-decoupled' ^{1}H spectrum as described before.[27] Projected spectra from samples taken at different time points were assembled and analysed by PCA in the usual way.

PCA of all samples together gave a rather complicated trajectory in the PC1/PC2 plane but analysis of samples at each time point together with the zero time controls gave in most cases a separation on PC1 which could be simply interpreted.[46] Relative to controls a consistent increase was seen in treated plants at most time points for flavonoid glycosides, sinigrin, Val, Thr, Ala and a consistent decrease of malic acid, carbohydrates (Glc and Suc) and Gln. Other compounds (fumaric acid, sinapoyl-malate) apparently showed an initial increase followed by a decline. Although the results from this type of study are largely descriptive and neither comprehensive nor quantitative enough for detailed correlation with results of microarray or proteomics experiments the NMR technique does have the benefit of giving an overall picture of the changes occurring, not confined to one class of metabolites.

Two studies have employed NMR profiling to study changes in the *Catharanthus* roseus (periwinkle) leaf metabolome following infection with phytoplasma. In the first of these papers *C. roseus* was infected with 10 types of phytoplasma.[47] Plant material was extracted with a methanol/water/chloroform mixture and then divided into organic and aqueous fractions. The major signals identified in the chloroform fraction for both healthy and infected plants were from the alkaloid vindoline. The aqueous fraction gave richer spectra and showed greater differences between healthy and infected leaves. PCA and measurement of the intensities of signals shown to be significant in the PC loadings indicated that there were increased levels in infected leaves of vindoline, loganic acid and secologanin (the last two compounds, identified in the aq. extract are precursors of vindoline), chlorogenic acid and other polyphenols, Glc and Suc.

The second paper used ^{1}H NMR to measure mono- and disaccharides (mainly Glc, Fru and Suc) and acids (lactic, malic) in *C. roseus* leaves following infection with *Spiroplasma citri* phytoplasma.[48] Plants inoculated with *S. citri* wild type and a mutant unable to take up Glc suffered impaired growth. They were analysed 5 weeks after infection and showed increased Glc and Suc accumulation in *C. roseus* leaves with increased Fru in the case of the mutant. A second mutant, unable to take up Fru, was non-pathogenic and showed the same pattern of sugar composition as healthy plants. The results showed that *S. citri* utilises Fru preferentially over Glc and allowed a model to be proposed of how the balance of sugar concentrations in the mature leaf is perturbed by the phytoplasma.

Metabolic changes in *Nicotania tabacum* (tobacco) leaves following infection with tobacco mosaic virus have been investigated using ^{1}H NMR.[49] Changes were followed with time (1, 3, 7, 10 days post-inoculation) for both locally infected leaves and 'systemically acquired resistance' (SAR) leaves in the same plants. A complex picture emerged with some changes common to both inoculated and SAR leaves and others observed only for one type (*e.g.* decrease of inositol in inoculated leaves only: accumulation of cembranoids and related compounds in SAR leaves). Changes discussed in some detail include increases in levels of 5-*O*-caffeoyl quinic acid and compounds containing α-linolenic acid chains and the decrease of inositol.

New Technologies and Hyphenated Techniques

Two main goals driving innovations in NMR technology have always been the improvement of sensitivity and the reduction of signal overlap. Two types of application can be distinguished: analysis of complex mixtures where the goal is to increase the number of quantifiable compounds without a prior separation step; and the high throughput acquisition of NMR and other data needed for structure determination of unknowns using the smallest possible amount of purified compound after an on-line or off-line chromatographic separation.

In a proof of principle paper[50] it was shown that 2D HSQC spectra (^{1}H–^{13}C or ^{1}H–^{15}N correlation) could be obtained in a reasonable time from *Arabidopsis* seedling extracts or even '*in vivo*' from hydrated seeds following universal ^{13}C or ^{15}N labelling. One advantage of the method is that the much greater dispersion of ^{13}C (or ^{15}N) chemical shifts in comparison with ^{1}H means that more metabolites can potentially be cleanly quantified *via* cross-peak intensities than is the case with conventional 1D ^{1}H spectra. This is not yet a routine high throughput method since, apart from the plant growth requirements, each 2D experiment would take about 200× longer than a conventional 1D experiment. However the authors calculate that with a system currently under construction (900 MHz spectrometer equipped with a cryoprobe) the acquisition time could be reduced to 5 min from the current 5 hr (on a 500 MHz spectrometer with no cryoprobe) and furthermore that the number of metabolites detected could be increased from the current 100–200 to 500–1000.

Screening of plant extracts for natural products with biological activity is one of the most actively pursued routes to novel drug discovery. There is also much interest in 'non-nutrient' but biologically active components of plant-based foods. Complete characterisation of the often complex structures of plant secondary metabolites can usually not be achieved with MS data alone but needs 1D and 2D NMR data as well. The development of methods for 'on-line' LC/NMR

as a complement to LC/MS has seen great progress over the last 20 years or so.[51] Much of it has been devoted to overcoming the limitations of real time 'on-line' NMR through the development of probes with high sensitivity flow cells (active volume 60–120 μL), loop storage systems that permit extended NMR accumulation times following transfer of the 'peak' to the measurement cell, special solvent suppression and decoupling strategies, and automation of the whole chromatography/peak selection/NMR data acquisition sequence.

In one of the first plant related experiments that used an integrated LC/UV/NMR/MS system a number of quercetin and phloretin glycosides from apple peel extracts were selected on the basis of their UV or MS data, then each peak was transferred from a loop storage system under automation for measurement of conventional ^{1}H NMR and 1D TOCSY spectra.[52] It was possible to determine the types of sugar ring present in each compound (in some cases where the MS was ambiguous) and to determine the linkage position of the sugars to the aglycone.

More recently automated solid phase extraction (SPE) units have been introduced to trap compounds as they elute from the HPLC column.[53] Provided that the compound is efficiently trapped by the SPE cartridge this system offers several advantages over the previous 'passive' loop storage method. After drying of the mixed (protonated) organic/aqueous solvent phase the compound is eluted from the cartridge and into the NMR probe with a pure deuterated solvent (*e.g.* CD_3CN or CD_3OD) using a volume of solvent (typically 30 μL) that matches the volume of the flow cell. This serves both to concentrate the compound and to standardise and simplify the solvent suppression requirements in comparison with the loop storage method. The SPE unit also allows for multiple trapping through repeat chromatography runs if necessary, *e.g.* in order to collect enough material for 2D heteronuclear spectroscopy. A further sensitivity gain (up to 4×) may be obtained by using a flow probe in which both the NMR coil and the preamplifier are cryo-cooled.

Use of a complete system of this type was first illustrated for identification of diverse compounds (flavonoids, phenolic acids, a monoterpene) present in an oregano extract.[53] There have been several subsequent reports on the use of similar systems to characterise natural products of interest in medicinal plants,[54,55] usually applying the SPE-NMR method as a final stage after a series of prior fractionation/concentration stages. Some of the potential of the SPE method can be seen from a comprehensive investigation of phenolic compounds present in polar extracts of virgin olive oils.[56] These samples gave complex chromatograms (UV detection) from which 27 compounds could be trapped by SPE. The NMR spectra obtained were of high enough quality to allow all these compounds to be identified, including several that had not been previously recognised in olive oil.

A somewhat different technological approach has been taken to high throughput screening of plant extracts and building of natural product libraries.[57] It employs a sequence of fractionation (fractions contain up to 5 compounds), biological activity screening to identify active fractions, further chromatography and screening to identify active compounds and characterisation of those compounds by MS and NMR. Compounds of interest are isolated (5–50 μg) and NMR spectra obtained by syringe infusion into an NMR probe fitted with a capillary microcoil (compound dissolved in 3 μL CD_3OD, active volume of the coil is 1.5 μL) that gives optimal sensitivity for mass limited samples and minimises interference from solvent and associated impurity signals. On a 600 MHz instrument 5 μg of compound is enough to carry out 1D and COSY ^{1}H NMR experiments and 50 μg is sufficient for 2D heteronuclear studies.

It is now also possible to measure very small volume samples (such as those obtained by elution from SPE cartridges) in tubes with outside diameter down to 1 mm. High sensitivity is obtained by using a probe fitted with a coil which matches the tube diameter. Although LC/NMR is still much less widely used than LC/MS the cumulative improvements in sensitivity achieved through the above modifications make it much more attractive than previously. The availability of high throughput LC/NMR strategies is particularly important in metabolomics since 'non-targeted' profiling studies, especially those that use LC/MS are revealing the existence of many unknown compounds and NMR data will be essential for a full characterisation of these unknowns.

Conclusions

Routine treatment of NMR data sets with multivariate statistics has now become accessible to all practitioners with the wide availability of software packages from instrument manufacturers and specialist companies. Probably the most troublesome remaining problem in the area of data analysis is caused by changes of chemical shift that result from inter-sample differences in pH and ionic composition.[58] Although there are continuing effects to resolve this data registration problem[59] there is no readily available general solution yet.

Some newly developed statistical methods that have been applied to NMR of biofluids will no doubt have an impact in plant metabolomics also. These include methods that simplify the interpretation of loadings plots (OPLS-DA[60]), the identification of unknowns from NMR data alone (statistical total correlation spectroscopy, STOCSY[61]) or NMR and LC/MS data combined (statistical heterospectroscopy, SHY[62]). Supervised methods such as genetic programming[63] and genetic algorithms[64] are able to pick out small groups of chemical shift values that are most effective for differentiating between samples from different groups and may identify minor signals that would be difficult to locate from standard PCA or PLS loadings plots.

Confirming the chemical origin of unknown signals arising from loadings plots or feature selection can be very time consuming, even though there is a high probability that the

required information is already to be found somewhere in the literature. There are now a number of freely available searchable electronic data bases, such as HMDB for human metabolites and SDBS for general organic chemicals that include assigned NMR spectra. Although the task would be a huge one (and would have to be collaborative, possibly with contributions from experts in particular classes of compound) there is no doubt that a similar NMR data base of plant derived compounds would be beneficial to the progress of plant metabolomics.

The advantages and limitations of NMR for metabolomics (especially in comparison with mass spectrometry) were mentioned in the Introduction. It does not provide an explicit quantity for each individual compound in the mixture (as the MS methods do) which makes it less useful than MS for some of the more theoretical treatments of metabolomics data. However new technology has led to notable improvements in NMR sensitivity and from the practical point of view modern equipment makes NMR profiling highly robust and quantitatively reliable. For these reasons it will probably find its greatest use in future in a high throughput screening and diagnostic role, allowing large numbers of samples to be surveyed and pointing the way towards more detailed investigations by complementary methods.

References

1) O. Fiehn: *Comp. Funct. Genomics* **2**, 155–168 (2001).
2) W. B. Dunn, N. J. C. Bailey and H. E. Johnson: *Analyst* **130**, 606–625 (2005).
3) S. Rochfort: *J. Nat. Prod.* **68**, 1813–1820 (2005).
4) P. Krishnan, N. J. Kruger and R. G. Ratcliffe: *J. Exp. Bot.* **56**, 255–265 (2005).
5) R. G. Ratcliffe and Y. Shachar-Hill: *Biol. Rev.* **80**, 27–43 (2005).
6) J. Lewis, J. M. Baker, M. H. Beale and J. L. Ward: "Genomics for Biosafety in Plant Biotechnology," ed. by J. P. H. Nap, A. Atanassov and W. J. Stiekema, IOS Press, pp. 47–57, 2004.
7) G. Le Gall, S. B. Metzdorff, J. Pedersen, R. N. Bennett and I. J. Colquhoun: *Metabolomics* **1**, 181–198 (2005).
8) G. Le Gall, I. J. Colquhoun, A. L. Davis, G. J. Collins and M. E. Verhoeyen: *J. Agric. Food Chem.* **51**, 2447–2456 (2003). (Correction **52**, 3210 (2004)).
9) A. Gil, I. F. Duarte, I. Delgadillo, I. J. Colquhoun, F. Casuscelli, E. Humpfer and M. Spraul: *J. Agric. Food Chem.* **48**, 1524–1536 (2000).
10) G. Le Gall, M. Puaud and I. J. Colquhoun: *J. Agric. Food Chem.* **49**, 580–588 (2001).
11) A. Lommen, J. M. Weseman, G. O. Smith and H. P. J. M. Noteborn: *Biodegradation* **9**, 513–525 (1998).
12) G. Le Gall, I. J. Colquhoun, A. L. Davis, G. J. Collins and M. E. Verhoeyen: *J. Agric. Food Chem.* **51**, 2447–2456 (2003).
13) J. L. Ward, C. Harris, J. Lewis and M. H. Beale: *Phytochemistry* **63**, 949–957 (2003).
14) Y. H. Choi, S. Sertic, H. K. Kim, E. G. Wilson, F. Michopoulos, A. W. M. Lefeber, C. Erkelens, S. D. Prat Kricun and R. Verpoorte: *J. Agric. Food Chem.* **53**, 1237–1245 (2005).
15) A. Moing, M. Maucourt, C. Renaud, M. Gaudilllère, R. Brouquisse, B. Lebouteiller, A. Gousset-Dupont, J. Vidal, D. Granot, B. Denoyes-Rothan, E. Lercetau-Köhler and D. Rolin: *Func. Plant Biol.* **31**, 889–902 (2004).
16) H. P. J. M. Noteborn, A. Lommen, R. C. van der Jagt and J. M. Weseman: *J. Biotech.* **77**, 103–114 (2000).
17) P. S. Belton, I. Delgadillo, A. M. Gil, P. Roma, F. Casuscelli, I. J. Colquhoun, M. J. Dennis and M. Spraul: *Magn. Reson. Chem.* **35**, S52–S60 (1997).
18) A. P. Sobolev, A. Segre and R. Lamanna: *Magn. Reson. Chem.* **41**, 237–245 (2003).
19) A. P. Sobolev, E. Brosio, R. Gianferri and A. L. Segre: *Magn. Reson. Chem.* **43**, 625–638 (2005).
20) E. K. Kemsley: "Discriminant Analysis and Class Modelling of Spectroscopic Data," J. Wiley, Chichester, UK, 1998.
21) J. C. Lindon, E. Holmes and J. K. Nicholson: *Prog. Nucl. Mag. Res. Spectrosc.* **39**, 1–40 (2001).
22) E. Holmes, P. J. D. Foxall, J. K. Nicholson, G. H. Neild, S. M. Brown, C. R. Beddell, B. C. Sweatman, E. Rahr, J. C. Lindon, M. Spraul and P. Neidig: *Anal. Chem.* **220**, 284–296 (1994).
23) L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold: "Multi- and Megavariate Data Analysis. Principles and Applications," Umetrics Academy, Umea, 2001.
24) G. Le Gall and I. J. Colquhoun: "Food Authenticity and Traceability," ed. by M. Lees, Woodhead Publishing Ltd., Cambridge, UK, pp. 131–155, 2003.
25) M. Frédérich, Y. H. Choi, L. Angenot, G. Harnischfeger, A. W. M. Lefeber and R. Verpoorte: *Phytochemistry* **65**, 1993–2001 (2004).
26) H. K. Kim, Y. H. Choi, C. Erkelens, A. W. M. Lefeber and R. Verpoorte: *Chem. Pharm. Bull.* **53**, 105–109 (2005).
27) H. Kirk, Y. H. Choi, H. K. Kim, R. Verpoorte and E. van der Meijden: *New Phytologist* **167**, 613–622 (2005).
28) G. Le Gall, I. J. Colquhoun and M. Defernez: *J. Agric. Food Chem.* **52**, 692–700 (2004).
29) N. J. C. Bailey, J. Sampson, P. J. Hylands, J. K. Nicholson and E. Holmes: *Planta Med.* **68**, 734–738 (2002).
30) Y. Wang, H. Tang, J. K. Nicholson, P. J. Hylands, J. Sampson, I. Whitcombe, C. G. Stewart, S. Caiger, I. Oru and E. Holmes: *Planta Med.* **70**, 250–255 (2004).
31) H. A. Kuiper, E. J. Kok and K.-H. Engel: *Curr. Opin. Biotechnol.* **14**, 238–243 (2003).
32) F. Cellini, A. Chesson, I. Colquhoun, A. Constable, H. V. Davies, K.-H. Engel, A. M. R. Gatehouse, S. Kärenlampi, E. J. Kok, J.-J. Leguay, S. Lehesranta, H. P. J. M. Noteborn, J. Pedersen and M. Smith: *Food Chem. Toxicol.* **42**, 1089–1125 (2004).
33) R. D. Firn and C. G. Jones: *Nature* **400**, 13–14 (1999).
34) G. Le Gall, M. S. DuPont, F. A. Mellon, A. L. Davis, G. J. Collins, M. E. Verhoeyen and I. J. Colquhoun: *J. Agric. Food Chem.* **51**, 2438–2446 (2003).
35) A. Charlton, T. Allnutt, S. Holmes, J. Chisholm, S. Bean, N. Ellis, P. Mullineaux and S. Oehlschlager: *Plant Biotech. J.* **2**, 27–35 (2004).
36) H.-K. Choi, Y. H. Choi, M. Verberne, A. W. M. Lefeber, C. Erkelens and R. Verpoorte: *Phytochemistry* **65**, 857–864 (2004).
37) C. Manetti, C. Bianchetti, M. Bizzarri, L. Casciani, C. Castro, G. D'Ascenzo, M. Delfini, M. E. Di Cocco, A. Laganà, A. Mic-

cheli, M. Motto and F. Conti: *Phytochemistry* **65**, 3187–3198 (2004).
38) S. J. W. Hole, P. W. A. Howe, P. D. Stanley and S. T. Hadfield: *J. Biomol. Screening* **5**, 335–342 (2000).
39) N. Aranibar, B. K. Singh, G. W. Stockton and K.-H. Ott: *Biochem. Biophys. Res. Comm.* **286**, 150–155 (2001).
40) K.-H. Ott, N. Aranibar, B. Singh and G. W. Stockton: *Phytochemistry* **62**, 971–985 (2003).
41) K. A. Aliferis and M. Chrysayi-Tokousbalides: *J. Agric. Food Chem.* **54**, 1687–1692 (2006).
42) T. W.-M. Fan, A. N. Lane, J. Pedler, D. Crowley and R. M. Higashi: *Anal. Biochem.* **251**, 57–68 (1997).
43) T. W.-M. Fan, A. N. Lane, M. Shenker, J. P. Bartley, D. Crowley and R. M. Higashi: *Phytochemistry* **57**, 209–221 (2001).
44) N. J. C. Bailey, M. Oven, E. Holmes, J. K. Nicholson and M. H. Zenk: *Phytochemistry* **62**, 851–858 (2003).
45) Y.-S. Liang, H. K. Kim, A. W. M. Lefeber, C. Erkelens, Y. H. Choi and R. Verpoorte: *J. Chromatogr. A* **1112**, 148–155 (2006).
46) O. Hendrawati, Q. Yao, H. K. Kim, H. J. M. Linthorst, C. Erkelens, A. W. M. Lefeber, Y. H. Choi and R. Verpoorte: *Plant Sci.* **170**, 1118–1124 (2006).
47) Y. H. Choi, E. C. Tapias, H. K. Kim, A. W. M. Lefeber, C. Erkelens, J. Th. J. Verhoeven, J. Brzin, J. Zel and R. Verpoorte: *Plant Physiol.* **135**, 2398–2410 (2004).
48) A. André, M. Maucourt, A. Moing, D. Rolin and J. Renaudin: *Mol. Plant-Microbe Interactions* **18**, 33–42 (2005).
49) Y. H. Choi, H. K. Kim, H. J. M. Linthorst, J. G. Hollander, A. W. M. Lefeber, C. Erkelens, J.-M. Nuzillard and R. Verpoorte: *J. Nat. Prod.* **69**, 742–748 (2006).
50) J. Kikuchi, K. Shinozaki and T. Hirayama: *Plant Cell Physiol.* **45**, 1099–1104 (2004).
51) K. Albert (ed.): "On-line LC-NMR and Related Techniques," J. Wiley, Chichester, UK, 2002.
52) A. Lommen, M. Godejohann, D. P. Venema, P. C. H. Hollman and M. Spraul: *Anal. Chem.* **72**, 1793–1797 (2000).
53) V. Exarchou, M. Godejohann, T. A. van Beek, I. P. Gerothanassis and J. Vervoort: *Anal. Chem.* **75**, 6288–6294 (2003).
54) C. Seger, M. Godejohann, L.-H. Tseng, M. Spraul, A. Girtler, S. Sturm and H. Stuppner: *Anal. Chem.* **77**, 878–885 (2005).
55) C. Clarkson, D. Stærk, S. H. Hansen and J. W. Jaroszewski: *Anal. Chem.* **77**, 3547–3553 (2005).
56) S. Christophoridou, P. Dais, L.-H. Tseng and M. Spraul: *J. Agric. Food Chem.* **53**, 4667–4679 (2005).
57) G. R. Eldridge, H. C. Vervoort, C. M. Lee, P. A. Cremin, C. T. Williams, S. M. Hart, M. G. Goering, M. O'Neil-Johnson and L. Zeng: *Anal. Chem.* **74**, 3963–3971 (2002).
58) M. Defernez and I. J. Colquhoun: *Phytochemistry* **62**, 1009–1017 (2003).
59) J. Forshed, R. J. O. Torgrip, K. M. Åberg, B. Karlberg, J. Lindberg and S. P. Jacobsson: *J. Pharm. Biomed. Anal.* **38**, 824–832 (2005).
60) O. Cloarec, M. E. Dumas, J. Trygg, A. Craig, R. H. Barton, J. C. Lindon, J. K. Nicholson and E. Holmes: *Anal. Chem.* **77**, 517–526 (2005).
61) O. Cloarec, M. E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes and J. K. Nicholson: *Anal. Chem.* **77**, 1282–1289 (2005).
62) D. J. Crockford, E. Holmes, J. C. Lindon, R. S. Plumb, S. Zirah, S. J. Bruce, P. Rainville, C. L. Stumpf and J. K. Nicholson: *Anal. Chem.* **78**, 363–371 (2006).
63) R. A. Davis, A. J. Charlton, S. Oehlschlager and J. C. Wilson: *Chemo. Intell. Lab. Syst.* **81**, 50–59 (2006).
64) E. K. Kemsley, G. Le Gall, A. D. Watson, L. J. Harvey, H. S. Tapp, J. R. Dainty and I. J. Colquhoun: *Br. J. Nutr.* **98**, 1–14 (2007).