Available online at www.sciencedirect.com
SCIENCE DIRECT®
Phytochemistry 65 (2004) 3187–3198
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

NMR-based metabonomic study of transgenic maize

Cesare Manetti [a,*], Cristiano Bianchetti [a], Mariano Bizzarri [b], Lorena Casciani [a], Cecilia Castro [a], Giuseppe D'Ascenzo [a], Maurizio Delfini [a], Maria Enrica Di Cocco [a], Aldo Laganà [a], Alfredo Miccheli [a], Mario Motto [c], Filippo Conti [a]

[a] *Dipartimento di Chimica, Università degli Studi di Roma "La Sapienza", Piazzale Aldo Moro 5, I-00185 Rome, Italy*
[b] *Dipartimento di Medicina Sperimentale e Patologia, Università degli Studi di Roma "La Sapienza", Piazzale Aldo Moro 5, I-00185 Rome, Italy*
[c] *Istituto Sperimentale per la Cerealicoltura, Sez. Bergamo, Via Stezzano 24, I-24126 Bergamo, Italy*

Received 19 July 2004; received in revised form 13 October 2004

Abstract

The aim of this research was to verify the possibility of identifying and classifying maize seeds obtained from transgenic plants, in different classes according to the modification, on the basis of the concerted variation in metabolite levels detected by NMR spectra. It was possible to recognise the discriminant metabolites of transgenic samples as well as to classify non-a priori defined samples of maize. It is important to underline that the obtained results are useful to point out the metabolic consequences of a specific genic modification on a plant, without using a targeted analysis of the different metabolites, in fact it was possible to classify the seeds also without the complete assignment of the spectra. The analysis was performed by applying multivariate techniques (principal component analysis and partial least squares-discriminant analysis) to NMR data.

Keywords: Maize; Metabonomics; Assay for genetic modification; Nuclear magnetic resonance; Principal component analysis; Partial least squares-discriminant analysis

1. Introduction

Maize is a major crop plant of essential agronomic interest as well as a model plant to evaluate chemical, physical and environmental effects and for use in genetic studies. With the development of plant genetic engineering technology, many transgenic strains of this monocotyledonous plant have been produced over the past decade.

In particular, field-cultivated insect-resistant Bt-maize hybrids are the centre of an intense debate between proponents and organizations recalcitrant to genetically modified organisms (GMOs). This debate, which addresses both safety and ethical aspects, has raised questions about the impact of genetically modified crops on the biodiversity of traditional landraces and on the environment (Hails and Kinderlerer, 2003).

Among the many problems concerning the products obtained by genetic transformation, little attention has so far been paid to the effects induced by these transformations on the metabolic processes, which are not directly dependent on the transformation itself.

The following points are of fundamental interest:

(a) determination of the genetic modification in terms of metabolite level variations compared to the non-modified species (equivalence);
(b) possibility of recognising and classifying the products into classes corresponding to specific genic transformations.

* Corresponding author. Tel.: +39 06 49913058; fax +39 06 4455278.
E-mail address: c.manetti@caspur.it (C. Manetti).

0031-9422/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2004.10.015

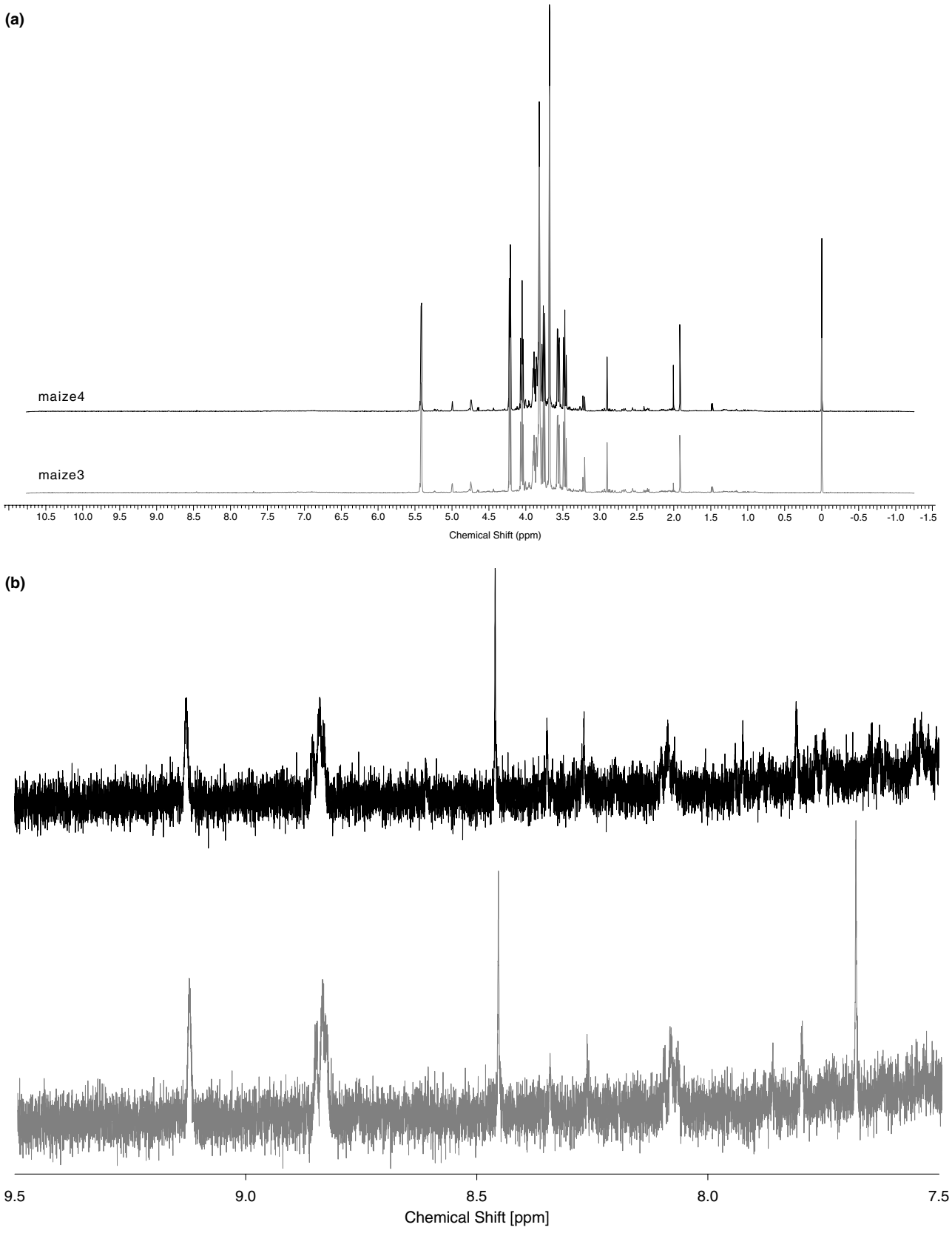

Fig. 1. (a) Spectra of the hydro-alcoholic extracts of seeds belonging to the groups maize3 and maize4. (b) An expansion of the spectra of the hydro-alcoholic extracts of seeds belonging to the groups maize3 and maize4, corresponding to the region 7.5–9.5 ppm.

A powerful tool in this respect is offered by metabonomics. Metabonomics (Lindon et al., 2001) represents an emerging holistic approach complementary to genomics and proteomics for studying the complex biological system response to chemical and physical input and also to genetic variations. The main purpose of the "-omics" technologies is the *non-targeted characterization* of all the genetic products (transcripts, proteins and metabolites) present in a specific biological system. Because of their characteristics, these technologies can afford global insight into the cell active processes, without any loss of intrinsic complexity.

Similar to transcriptomics and proteomics, determination of the biological system metabolites defines its metabolome, in other words its metabolic fingerprint, which allows us to identify and to dynamically follow its growth and/or its responses to environmental conditions.

Plants are sessile systems that are unable to escape environmental pressures. As a result, they have evolved a dazzling array of flexibility in their responses to environmental conditions such as light or dark, drought, temperature, nutritional supply, microbial invasion. Thus, the plant system comprises a genotype by environmental responses, producing a specific geno–phenotype relationship that is heavily dependent on the growth stage and several studies are performed to investigate this kind of "perturbation". In this way, a gene function must be defined in the system and environment context (Weckwerth, 2003).

In the last few years, genomics and proteomics have been used to identify genes and proteins revealing different expressions due to systemic perturbation, and also metabolite profiling have been monitored (Fiehn et al., 2000; Bailey et al., 2003; Defernez et al., 2004; Ward et al., 2003).

All things considered, when the aim of the study is to investigate the consequences of a specific genic modification on a plant, the hierarchical stream genomics, transcriptomics, proteomics and metabonomics loses all its meaning. The different steps must be considered as interdependent and integrated. Thus, the changes in the genic expression determine variations in transcription and in protein contents, which in turn determine variations in the metabolite levels and fluxes.

The metabolite network can influence the protein network by feedback inhibition and positive modulation and also interact with gene network through sensing and signal transduction. Clearly, determination of the network correlations is more important than the targeted determination of the single substance level.

One elective technique for metabonomics is Nuclear Magnetic Resonance (NMR) Spectroscopy, which allows us to obtain qualitative and quantitative data of many metabolites of the biological system as a whole, in a non-destructive way, without losing the complexity of the systems (Lindon, 2004).

Additionally, NMR spectra, analysed by multivariate analysis techniques, such as Principal Component Analysis (PCA) and Partial Least Squares-Discriminant Analysis (PLS-DA), allow us to evaluate the system in terms of the different metabolite levels variation and the metabolomic network behaviour in terms of covariance.

The aim of this research is to verify the possibility of an NMR metabonomic approach to identify and classify

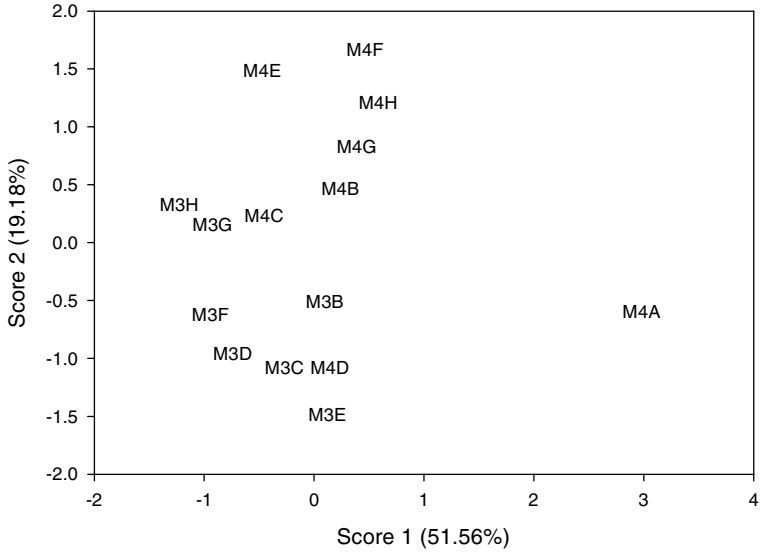

Fig. 2. Representation of seeds belonging to the groups maize3 and maize4 in the PCA space spanned by the score of the first and the second principal components. The sample M3A has not been considered in the chemometric analysis, because it is an outlier localised in a completely separated part of the space spanned by principal components 1 and 2.

maize seeds originating from transgenic plants, in different classes according to the modification, on the basis of the concerted variation in metabolite levels detected by NMR spectra.

2. Results and discussion

2.1. Application of metabonomic analysis to a transgenic–control couple

Transgenic seeds (belonging to the group maize4) and their unmodified controls (belonging to the group maize3) have been analysed. An example of the spectra of the hydro-alcoholic extracts of the seeds is reported in Fig. 1(a) and (b). First of all, we applied an unsupervised method, PCA, to operate an exploration of the data.

PCA, applied to the bucketed spectra of the samples belonging to the groups maize3 and maize4, gave a good representation of the data with 3 PC, that respectively accounted for 51.6%, 19.2% and 9.0% of the variance in the dataset, cumulatively equal to 79.8%.

The PCA scores obtained are shown in Fig. 2.

In order to obtain a predictive model that allows us to classify different groups of seeds belonging to the different groups, we took spectral variable correlations into consideration (in this specific case the bucketed spectra regions integrals of the transgenic and control seeds) and applied the supervised PLS-DA analysis technique. The results are summarised in Table 1. The model effects and dependent variables show how much predictor and response variation are explained by each PLS-DA factor.

A representation in the space spanned by the PLS-DA scores of the first and the second latent variables of seeds belonging to the groups maize3 and maize4 is shown in Fig. 3. It is evident that the first factor manages effectively to discriminate between transgenic and their unmodified control seeds.

Table 1
Results obtained from the cross validation of the PLS-DA model for the transgenic–unmodified couple maize3–maize4

Number of extracted factors	Model effects	Dependent variables	Root mean PRESS
1	46.5254	60.1777	0.795733
2	22.1018	25.1434	0.548787
3	8.1519	11.4666	0.397966
4	4.0703	2.2860	0.352462
5	6.9362	0.4712	0.356518
6	3.0092	0.3012	0.330572
7	1.3887	0.1350	0.330259
8	1.8350	0.0170	0.330501
9	1.1879	0.0016	0.330682
10	0.9335	0.0003	0.33027

The model effects and dependant variables show how much predictor and response variation are explained by each PLS-DA factor.

We therefore took the correlation structure into consideration: in Fig. 4(a), the 1st factor *X*-weights vs. 2nd factor *X*-weights are shown. By analysing the weights given to each of the original variables, i.e. the degree of correlation between the variables and the direction of the new model, it is possible to determine the hierarchic importance of the variables for the discrimination between the two groups (transgenic and control) of samples. A positive value in the loadings plot implies a positive correlation with the scores in the first latent variable. Thus, all variables with positive values are positively correlated with the samples with positive scores, whilst the variables with negative values are

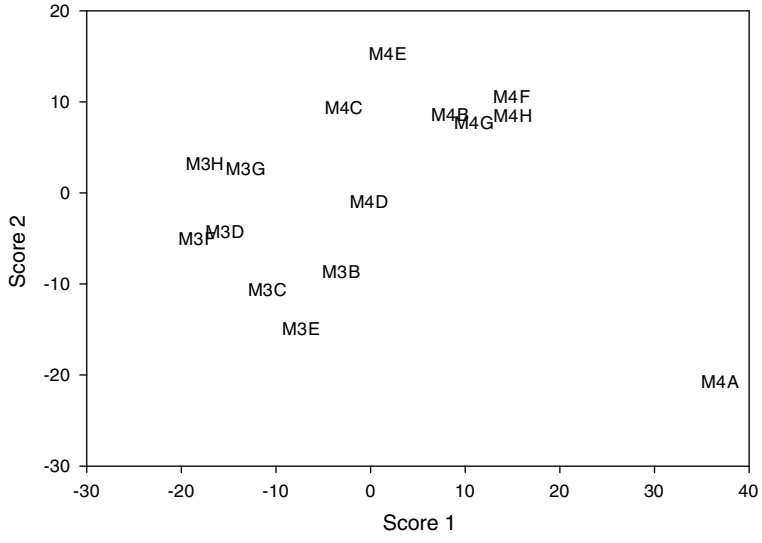

Fig. 3. Representation of seeds belonging to the groups maize3 and maize4 in the PLS-DA space spanned by the score of the first and the second latent variables. As already said for the PCA score plot, the sample M3A is an outlier and it has not been included in the analysis.

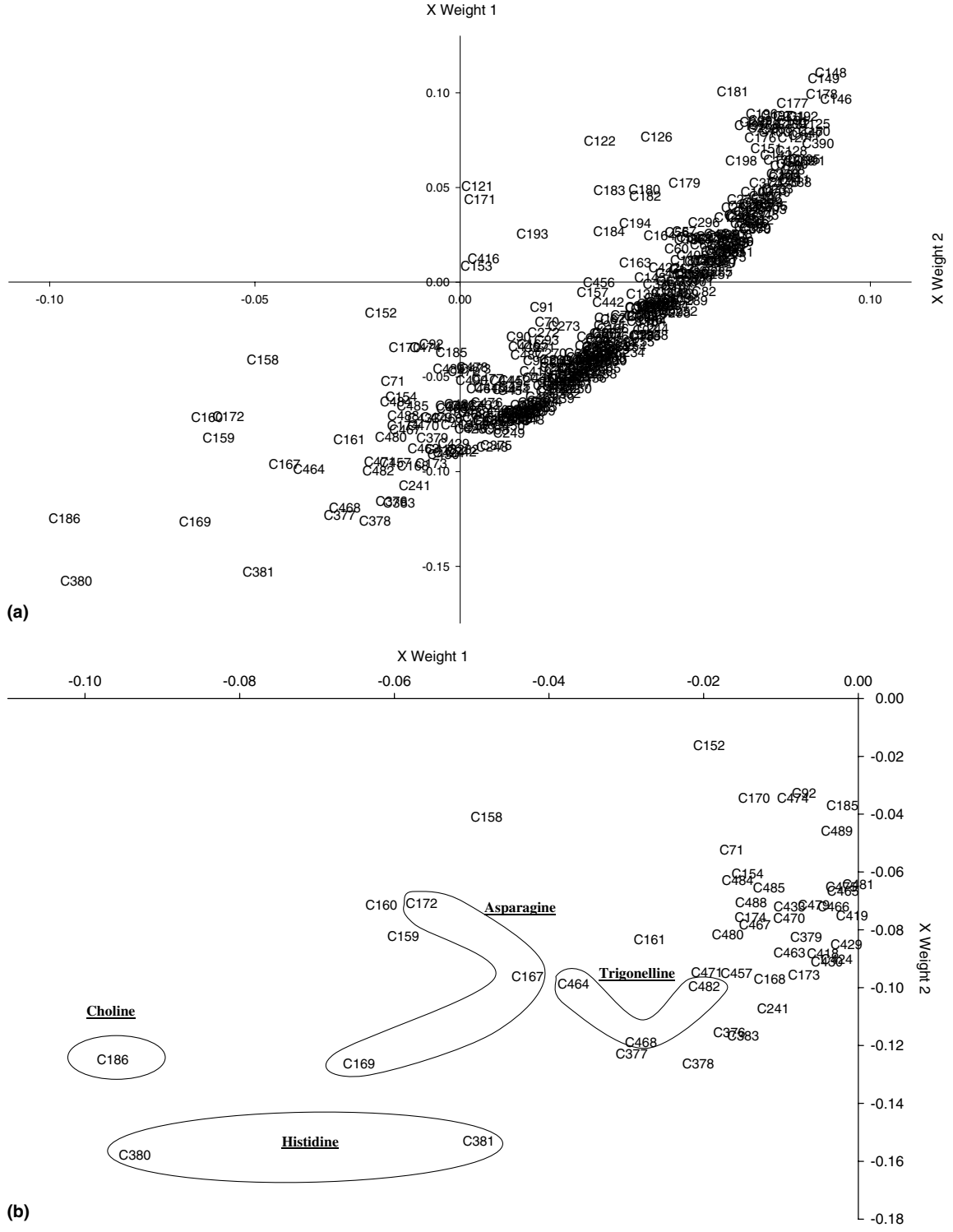

Fig. 4. (a) Plot of the first loading vs. the second one, obtained by PLS technique. A large negative value of loading relative to metabolites, indicates higher levels in the control samples (the ones that had negative scores), and lower levels in the transgenic samples (the ones that had positive scores). (b) An enlargement of Fig. 6(a), showing, in particular, the third quadrant and enlighting the most influential variables in discriminating between the two maize groups.

correlated with the samples with negative scores, making it possible to identify the metabolite that discriminates between the two groups (Bailey et al., 2003). A large negative value of loading relative to metabolites, in Fig. 4(a), indicates higher levels in the control samples (the ones that had negative scores), and lower

levels in the transgenic samples (the ones that had positive scores).

As an example of the interpretation of the graph showed, looking to the third quadrant (Fig. 4(b)) we can identify the variables corresponding to specific metabolites which were lower in transgenic samples than in controls: choline (*singlet* 3.21 ppm), asparagine (*double doublet* 2.86 ppm; *double doublet* 2.96 ppm), histidine (*singlet* 7.07 ppm) and trigonelline (*singlet* 9.13 ppm; *multiplet* 8.84 ppm) (the assignment of each substance

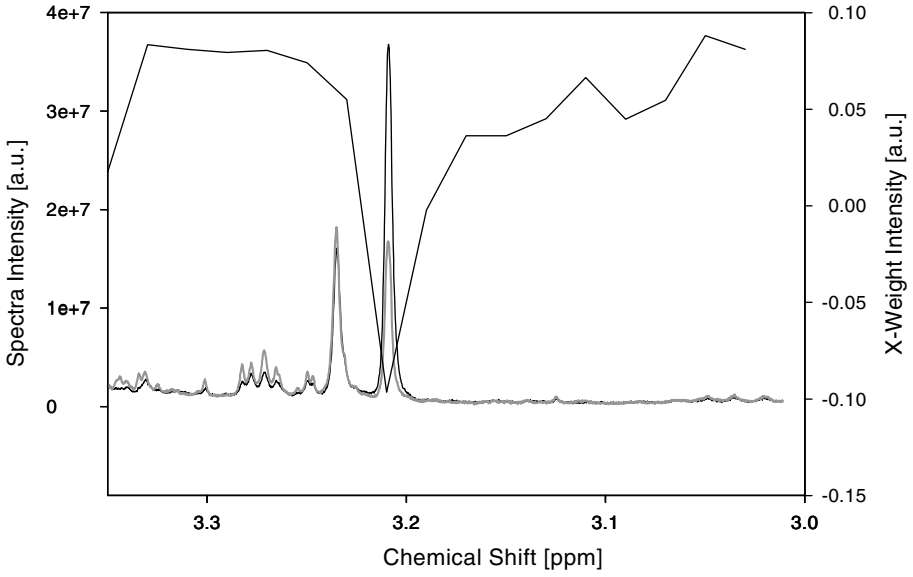

Fig. 5. The choline region in maize3 (in black) and maize4 (in gray) spectra and in the first latent variable.

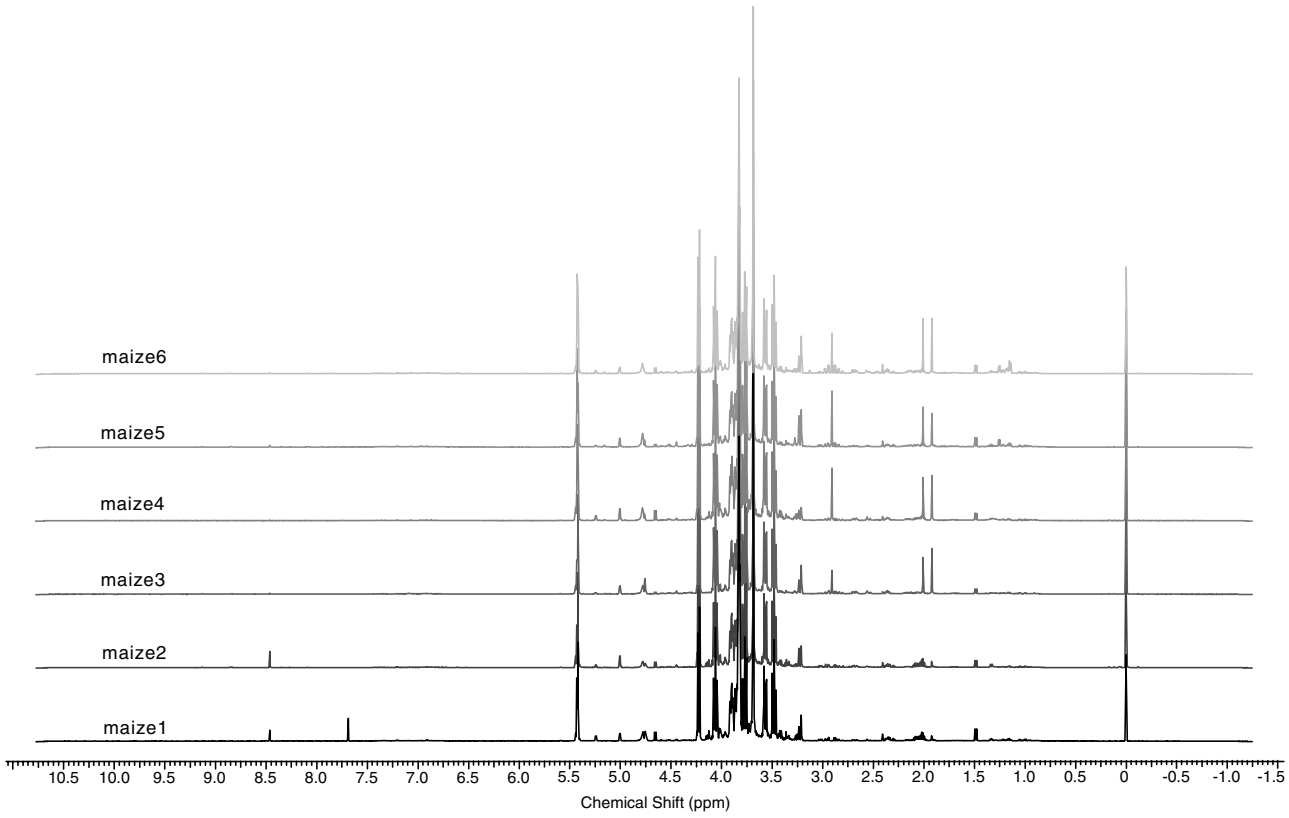

Fig. 6. Spectra of the hydro-alcoholic extracts of seeds, one for each maize group.

was based on comparison with the chemical shift values reported in literature (Sobolev et al., 2003; Defernez et al., 2004; Shachar-Hill et al., 1996)).

By way of example, in Fig. 5 the choline regions in maize3, maize4 spectra and in the 1st factor *X*-weight indicating the signal that contribute to discrimination are shown.

In this way, elucidate the signals that are key in the separation between the two samples groups has been directly obtained, without pre-analysis derivatization and thus pre-selection of the expected metabolites.

2.2. Application of PLS to classify unknown samples of maize

We subsequently considered the possibility of classifying maize seeds, whose genetic modification is not defined a priori in the method.

This allows us to assert the possibility to construct an NMR spectra database in order to define different maize metabolome, as Nicholson's group has made for biofluids (Holmes and Antti, 2002).

On the basis of the NMR spectra, it was possible to build up a PLS-DA model. In Fig. 6, hydro-alcoholic extracts spectra are shown, one for each maize group.

We thus took six maize groups into consideration (three different genetic modifications and their own controls), each formed by 8 samples. In this case, the input *Y* "dummy" matrix has a number of columns equal to the number of groups. To each sample, there is a row containing 1 in columns corresponding to the right group and zero in all the others.

To choose the model with the right number of latent variables, a cross validation technique was used and the results are reported in Table 2. As a result, a model with 13 latent variables was chosen, as it corresponds to the minimum predicted residual sum of squares (PRESS) value.

This model was used to evaluate its predictive capabilities for other 12 different seed samples The output of PLS-DA procedure gives a *Y* matrix that can be used to classify unknown samples. Indeed, looking at this matrix the sample is assigned to the group that shows the maximum *y* value. In Table 3, *y* values for this second set of 12 different samples are displayed.

Table 2
Results obtained from the cross validation of the PLS-DA model used to classify unknown maize samples

Number of extracted factors	Model effects	Dependent variables	Root mean PRESS
1	41.5198	6.7402	1.05604
2	11.8754	9.8096	1.087884
3	14.1704	7.8759	1.0745
4	5.1089	15.4728	1.020006
5	4.2491	12.9411	0.969842
6	2.9438	13.5194	0.883287
7	2.7672	8.6755	0.906072
8	1.9184	8.5668	0.83381
9	1.7871	3.6077	0.853818
10	1.5014	3.6685	0.762747
11	1.3183	1.1916	0.759206
12	1.1748	1.4480	0.736652
13	*0.9289*	*0.8905*	*0.726226*
14	1.0991	0.7913	0.728478
15	0.5772	1.2022	0.751386
16	0.5770	0.7262	0.767855

Applying PLS-DA it was possible to obtain the correct classification of these samples in three different groups, each composed of four seeds, corresponding to maize1, maize3, maize5.

The results showed that the dimension reduction, generated by multivariate analysis of NMR spectra is based upon the existence of correlations between the original variables, is effective. In the case of completely independent variables no truncation of dimension could be possible. This implies that a successful analysis (i.e., the reaching of a relevant portion of variance explained) corresponds to a given number of original variables (metabolites) highly loaded on the first few components. This is a result directly emerging from the data without any subjective interpretation (Giuliani et al., 2004).

Table 3
Y matrix obtained for the second set, formed by 12 samples, not used to construct the model

Sample	Group1	Group2	Group3	Group4	Group5	Group6
Maize1I	*0.777006*	0.105033	0.238784	−0.3067	−0.01416	0.200038
Maize1L	*0.822596*	0.364114	−0.00673	−0.05281	−0.43344	0.306276
Maize1M	*0.605178*	−0.14805	0.209341	0.198064	−0.10013	0.2356
Maize1N	*0.407808*	−0.0317	0.225597	0.133936	0.177597	0.086761
Maize3I	0.011782	0.036883	*0.897644*	−0.00734	−0.01609	0.077114
Maize3L	−0.01163	−0.19545	*0.883801*	0.118117	−0.08281	0.287965
Maize3M	−0.08118	−0.06844	*0.923928*	0.123475	0.072519	0.029693
Maize3N	−0.11705	0.117507	*0.859392*	0.156886	0.120912	−0.13765
Maize5I	−0.18455	0.253777	0.259816	0.279073	*0.555209*	−0.16332
Maize5L	−0.14411	0.227277	0.264199	0.161204	*0.58734*	−0.09591
Maize5M	−0.12811	0.222366	0.321713	0.141874	*0.611829*	−0.16967
Maize5N	−0.1625	0.228232	0.287785	0.15676	*0.59786*	−0.10813

PLS-DA applied to the entire dataset shows that 13 variables are sufficient to explain almost 90% of total variability. To go in deep, in next future after a complete assignment of the spectrum, we will apply directly on the resulted metabolites concentrations the multivariate approach to investigate the metabolic network.

However, this second step is independent and not necessary, when the aim is only the classification of different seeds samples.

From these results, it appears useful to construct a NMR spectra database to evaluate the influence of environmental, chemical, physical and genetic input on the system.

Note that, for this purpose, it is essential to converge on a standardised sample preparation, in particular in terms of sample preparation.

3. Conclusions

Metabonomics represents an emerging holistic approach complementary to genomics and proteomics for studying complex biological system behaviour.

In a biological system, the metabolites can be considered the actual phenotypic expression, as the genic function can be viewed as the result of interconnected, non-hierarchical, regulatory processes in a gene–transcript–protein–metabolite–metabolic network. The metabolites concentrations in a cell system can be considered as the last response to the chemical, physical environmental as well as genetic changes.

In this article, we demonstrate that it is possible to evaluate the possible determination of the genetic modification in terms of comparison of the present metabolites with non-modified specimens of the same species (equivalence); to identify and classify maize seeds originating from transgenic plants on the basis of analysis using multivariate techniques of the ^{1}H NMR spectra (global system descriptors), which reveals correlations among the different metabolites, without using a targeted analysis of the different metabolites.

4. Experimental

4.1. Plant material

The seed samples used in this study (Table 4) were derived from the maize inbred lines La73 (maize3) and La17 (maize5) and their transgenic versions (respectively, maize4 and maize6) containing the *cry1Ab* gene (MON 810) from *Bacillus thuringensis*, conferring resistance to the European corn borer. MON810 was developed and kindly provided by Monsanto Co. (St. Louis, Mo). In addition, a seed sample of the B73 inbred lines (maize1) and its trangenic version (maize2) containing a modified *ZmRpd3-101* maize gene (Rossi et al., 1998), was included. For this last event the cDNA clone *ZmRpd3-101* containing the entire region of the *ZmRpd3* gene in an antisense orientation was transcriptionally fused to the 35 S CaMV constitutive promoter and to a T-DNA Nos terminator, using standard recombinant DNA techniques. This chimaeric gene was inserted into the pSC1 expression vector carrying the *ubi1-bar* sequence as selectable marker. The new clone, denoted as pRpd3-5.3 was used to transform maize plants. These transgenic maize events were created by polyethylene glycol-mediated direct-DNA uptake transformation of protoplasts derived from a suspension culture (Morocz et al., 1990.). Transgenic cells were selected on medium containing glufosinate (Basta resistence) and plants were regenerated as described by Morocz et al. (1990). The transformed plants were converted to the B73 background by backcrossing four times, followed by two self-pollinations. These transgenic lines were selected following RT–PCR analysis.

Table 4
List of the seed samples used in this study

Sample	Inbred line
Maize1	C4
Maize2	33 Homozygote AS ZmRpd3/101
Maize3	G03-1220 B73+
Maize4	03-1216 B73 Bt
Maize5	G03-1220 Mo17
Maize6	03-1218 Mo17 Bt

Plants of inbred lines La73, La17, and B73 and their transgenic versions were grown under greenhouse conditions at 25:18 °C (day:night) with a 16:8 (light:dark) hour cycle. At flowering, plants were self-pollinated; the ears were harvested after physiological maturity, dried at 30 °C and stored in sealed plastic bags at 4 °C. For each genotype, a seed sample derived from the central portion of a single ear was used for chemometric analyses.

4.2. NMR methods

4.2.1. NMR sample preparation

For each sample a single maize seed was weighed (200 mg ca.) and then frozen in a stainless steel mortar by liquid N_2, before being pulverised to a fine powder with a pestle chilled in liquid N_2 and maintained in liquid N_2 bath during the pulverization procedure.

Three ml of methanol/chloroform mixture (2:1) were added to the powder. The powder was stirred and 1 ml of chloroform and 1.2 ml of water were added (Bligh–Dyer modified) (Miccheli et al., 1988; Ricciolini et al., 1994). The sample was stored at 4 °C for 1 h and then centrifuged at 10,000 *g* for 20 min at 4 °C. The resulting upper hydro-alcoholic and lower chloroformic phases were separated. The extraction procedure was performed twice on the pellet in order to obtain a quantita-

tive extraction. After the second extraction, the two hydro-alcoholic phases obtained were re-collected, dried under N_2 flux, and stored at −80 °C prior to analysis.

4.2.2. NMR data collection

For the NMR spectra, the dried sample was dissolved in 1 ml of 0.5 mM TSP solution in D_2O PBS buffer (pH 7.4) to avoid chemical-shift changes due to pH variation. (Defernez and Colquhoun, 2003). The dissolved extracts were transferred to a 5-mm NMR tube.

NMR spectra were recorded on a Bruker (Bruker GmbH, Rheinstetten, Germany) DRX 500 spectrometer, operating at ^{1}H frequency of 500.13 MHz. ^{1}H NMR spectra were obtained at $T = 300$ K, 256 scans were acquired, with data collected into 64 k datapoints, and a spectral width of 12 ppm, using a 20-s delay for a full relaxation condition. The water resonance was suppressed by irradiation during 5-s relaxation delay at a power level 70 dB below max transmitter power setting (Rahman, 1989). Prior to Fourier transformation, an exponential multiplication was performed, using a line broadening equal to 0.09 Hz: this value represent an optimum balance between the noise reduction and the line broadening effects, considered digital resolution. Spectra were referenced to TSP (sodium salt of 3-(trimethylsilyl)propionic-2,2,3,3-d_4 acid) at a final concentration of 0.5 mM. TSP was used as a reference both for chemical shift (0.00 ppm) and quantitation of the signals (Defernez and Colquhoun, 2003).

The processing of the spectra was carried out using ACD Software. The spectra were phased, baseline corrected using the usual ACD routine selecting only two points located at the extremes of the spectra in the part that contains only noise. The spectra were scaled fixing the area of TSP signal to a value of 10.

4.2.3. NMR data pre-processing treatment

One-dimensional 500 MHz ^{1}H spectra were reduced to 499 discrete chemical shift regions by digitization to produce a matrix of sequentially integrated regions of 0.02 ppm in width between −0.5 and 9.5 ppm, using ACD/SpecManager 7.00 software (Advanced Chemistry Development Inc., 90 Adelaide Street West, Toronto, Ont., Canada M5H 3V9): column 1 corresponds to the bucket −0.5 to −0.48 ppm.

No region was excluded during the digitisation to standardise the procedure. This choice makes this step unsupervised and avoid the necessity of changing the considered regions at the occurrence of new signals in unknown samples. This characteristics is very useful when the purpose is the construction of a database.

Note that also the solvent region was included in the analysis: the water suppression procedure produces random variation on integrals of solvent region, so it results in uncorrelated noise that the multivariate approach, based on the eigenvectors of the covariance (or correlation) matrices (PCA and PLS-DA) filters out (Lebart et al., 1984; Broomhead and King, 1986).

Regions containing no signals or too overlapped regions were excluded only from the graphical representation of the analysis, to avoid an increase of the uninformative "noise".

4.3. Multivariate data analysis

4.3.1. Principal components analysis

This is a well-known multivariate technique, originally developed early last century (Spearman, 1904). This technique has had an almost universal application, ranging from hydrodynamics (Craddack, 1965; Preisendorfer, 1988; Ghil and Vantard, 1991) to sociological (Aitkin, 1974) and biological research (Gage et al., 1989; Giuliani et al., 1991).

The main purpose of PCA is to define the real dimensionality of the data field under study. When measuring N variables (NMR signals in our case) on K units (samples in our case), a situation is delineated which appears to be N-dimensional. However, these variables may be correlated in various ways among themselves, and so an equally satisfactory description could be obtained with a P ($P < N$) number of axis, which are called factors or components and represent the degree of freedom of the system.

From a geometrical point of view, these dimensions (factors or components or latent variables) represent the directions in the data field along which the variability of the data clouds is maximal (Lebart et al., 1984). From a mathematical point of view, components are eigenvectors of the correlation matrix among the original variables: they are orthogonal to each other and extracted by the algorithm in the order of percentage of explained variability. Thus, the first factor will be the one explaining the highest proportion of variation embedded in the original data matrix. Factors are constructed so as to have a mean value of zero and an unitary standard deviation over the entire dataset.

Different variations of PCA can be performed by varying the nature of the data in X. X can be mean-centred, or standardised (mean-centred and columns scaled to unit variance). One advantage of the first method is that the eigenvectors (or loadings) retain the scale of the original data, and will often resemble spectra. In contrast, the loadings obtained by standardised data are usually very unfamiliar in appearance. However, one advantage of this approach is that the PCA is influenced by all spectral features equally, whereas in the other approach, larger resonances tend to dominate. Consequently, the second data pre-processing method can be useful when minor constituents, with small spectral contributions are of primary interest (Belton et al., 1998). In our case, we applied this second data pre-processing method to be sure that

all the spectral regions make their contribution to the discrimination between the different samples.

4.3.2. Partial least square-discriminant analysis

PLS technique was originated by Herman Wold (1966) for the modelling of complicated datasets in terms of chains of matrices, the so-called path models. After this first use, PLS was applied to spectrometric calibration (Haaland and Thomas, 1988), to monitoring and controlling industrial processes (Wang et al., 2003) and in recent years to metabonomics (Lindon et al., 2001; Brindle et al., 2002).

PLS is a method for constructing predictive models when the factors are many and highly collinear. It maximises the covariance between the predictor space (matrix of NMR data, X) and the response space (matrix of the information on maize lines to which the seeds belong, Y). The overall goal is to use the factors to predict the responses in the population. This is achieved indirectly by extracting latent variables T (X-scores) and U (Y-scores) from sampled factors and responses, respectively. The extracted factors T are used to predict the U, and then the predicted Y-scores are used to construct predictions for the responses (Tobias, 1995).

In other words, the dataset is interpreted in terms of X- and Y-scores (T,U), X-loading (P), X- and Y-weights (w,c) and PLS regression coefficients (B) (Wold et al., 2001). Once a PLS model has been calculated and validated, it can be used for the prediction of class membership of unknown samples.

A particular version of this method (reported as PLS-DA) is done by a regression of the data (X) against a "dummy matrix" (Y), which describes variation according to class. In the training set, the input Y "dummy" matrix has a row, for each sample, containing 1 for the y variable corresponding to the right group and zero for all the others.

In Fig. 7, a schematic representation of the two procedures is shown.

4.3.3. Application of metabonomic analysis to a transgenic–control couple

We applied PCA and PLS-DA to a matrix containing pre-processed data relative to samples belonging to the

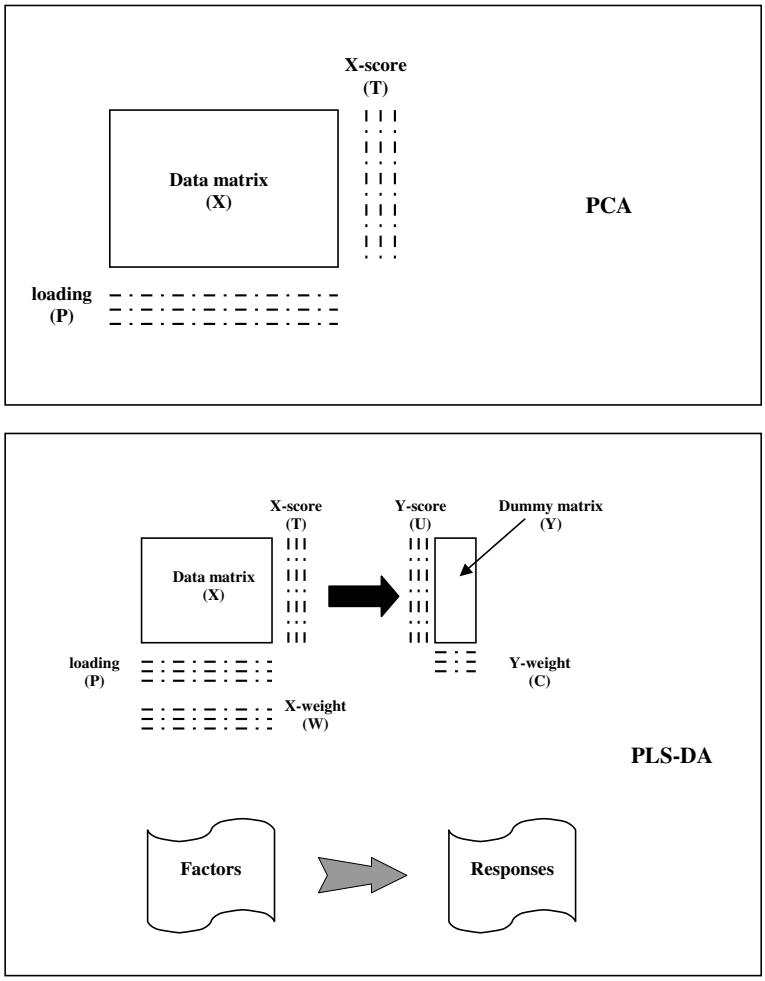

Fig. 7. A schematic representation of the two multivariate technique, PCA and PLS-DA.

groups maize3 and maize4 (each composed of 8 samples). In the PLS-DA procedure the Y matrix corresponds to a column containing an entry equal to zero or one, according to the seed line, for example maize3 or maize4. The SAS software (Statistical Advanced Software) v.8 (SAS Institute Inc., www.sas.com) was used for all the statistical analysis and the procedures are available from the authors on request (SAS Institute Inc., 1999).

4.3.4. Application of PLS-DA to classify unknown samples of maize

We constructed the PLS-DA model with a X matrix containing pre-processed data relative to samples belonging to six groups (each composed by 8 samples). The input Y "dummy" matrix has a number of columns equal to the number of groups. For each sample, the row contains 1 for the y variable corresponding to the right group and zero for all the others. The correct number of latent variables utilised to construct the model was chosen by cross validation. The most common technique is one-at-a time validation, unless the observed data is serially correlated, in which case either blocked or split-sample validation may be more appropriate. Note that one-at-a time validation is the most computationally intensive of the cross validation methods, since it requires a recomputation of PLS model for every input observation.

The number of factor chosen was the one that minimizes the PRESS. Its absolute minimum correspond to the correct number of extracted factors. PRESS is defined as sum of squared differences between predicted and observed y values (over all rounds)

$$\mathrm{PRESS} = \sum_i (y_i - \hat{y}_{\mathrm{in}})^2,$$

where $\hat{y}_{\mathrm{in}}$ is the prediction with model fitted with ith observation delated.

The validated model was used on a set, containing 12 new samples belonging to different groups. The new observation are not used in calculating the PLS-DA model, since they have no assigned values in the input Y matrix.

The output of the PLS-DA procedure gives a Y matrix with row also for the samples that can be used to classify these unknown samples. In fact, looking at this matrix the sample is assigned to the group that shows the maximum y value.

References

Aitkin, R., 1974. Mathematical Structure in Human Affairs. Heinemann Educational Books, London.

Atta-ur-Rahman, 1989. One and Two Dimensional NMR Spectroscopy. Elsevier, Amsterdam.

Bailey, N.J.C., Oven, M., Holmes, E., Nicholson, J.K., Zenk, M.H., 2003. Metabolomic analysis of the consequences of cadmium exposure in *Silene cucubalus* cell cultures via ^{1}H NMR spectroscopy and chemometrics. Phytochemistry 62, 851–858.

Belton, P.S., Colquhoun, I.J., Kemsley, E.K., Delgadillo, I., Roma, P., Dennis, M.J., Sharman, M., Holmes, E., Nicholson, J.K., Spraul, M., 1998. Application of chemometrics to the ^{1}H NMR spectra of apple juices: discrimination between apple varieties. Food Chem. 61, 207–213.

Brindle, J.T., Antti, H., Holmes, E., Tranter, G., Nicholson, J.K., Bethell, H.W., Clarke, S., Schofield, P.M., McKilligin, E., Mosedale, D.E., Grainger, D.J., 2002. Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics. Nat. Med. 8, 1439–1444.

Broomhead, D.S., King, G.P., 1986. Extracting qualitative dynamics from experimental data. Physica D 20, 217–236.

Craddack, J.M., 1965. A meteorological application of principal component analysis. Statistician 15, 143–156.

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

Defernez, M., Gunning, Y.M., Parr, A.J., Shepherd, L.V.T., Davies, H.V., Colquhoun, I.J., 2004. NMR and HPLC–UV profiling of potatoes with genetic modifications to metabolic pathways. J. Agric. Food Chem. 52, 6075–6085.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N., Wilmitzer, L., 2000. Metabolite profiling for plant functional genomics. Nat. Biotechnol. 18, 1157–1161.

Gage, F.H., Dunnett, S.B., Bjorklund, A., 1989. Age-related impairments in spatial memory are independent of those in sensory motor skills. Neurobiol. Aging 10, 347–352.

Ghil, M., Vantard, R., 1991. Interdecadal oscillations and the warming trends in global temperature time series. Nature 350, 324–327.

Giuliani, A., Capuani, G., Miccheli, A., Aureli, T., Ramacci, M.T., Conti, F., 1991. Multivariate data analysis in biochemistry: a new integrative approach to metabolic control in brain aging. Cell. Mol. Biol. 37, 631–638.

Giuliani, A., Zbilut, J.P., Conti, F., Manetti, C., Miccheli, A., 2004. Invariant features of metabolic networks: a data analysis application on scaling properties of biochemical pathways. Physica A 337, 157–170.

Haaland, D.M., Thomas, E.V., 1988. Partial least squares methods for spectral analysis. 1.Relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 60, 1193–1202.

Hails, R., Kinderlerer, J., 2003. The GM public debate:context and communication strategies. Nat. Rev. Genet. 4, 819–825.

Holmes, E., Antti, H., 2002. Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. Analyst 127, 1549–1557.

Lebart, L., Marineau, A., Warwick, K.M., 1984. Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices. Wiley, New York.

Lindon, J.C., Holmes, E., Nicholson, J.K., 2001. Pattern recognition methods and application in biomedical magnetic resonance. Prog. Nucl. Magn. Reson. Spectrosc. 39, 1–40.

Lindon, J.C., 2004. Metabonomics – Techniques and Applications. Business Briefing: Future Drug Discovery, 1–6 www.bbriefings.com/pdf/855/fdd041_metabometrix_tech.pdf .

Miccheli, A., Aureli, T., Delfini, M., Di Cocco, M.E., Viola, P., Gobetto, R., Conti, F., 1988. Study on influence of inactivation enzyme techniques and extraction procedures on cerebral phosphorylated metabolite levels by P-31 NMR spectroscopy. Cell. Mol. Biol. 34, 591–603.

Morocz, S., Donn, G., Nemeth, J., Dudits, D., 1990. An improved system to obtain fertile regenerants via maize protoplasts isolated

from a highly embryogenic suspension culture. Theor. Appl. Genet. 80, 721–726.

Preisendorfer, R.W., 1988. Principal Component Analysis in Metereology and Oceanography. Development in Atmospheric Science, vol. 17. Elsevier, Amsterdam.

Ricciolini, R., Miccheli, A., Di Cocco, M.E., Piccolella, E., Marino, A., Sammartino, M.P., Conti, F., 1994. Dexamethasone-dependent modulation of human lymphoblastoid B cell line through sphingosine production. Biochim. Biophys. Acta 1221, 103–108.

Rossi, V., Hartings, H., Motto, M., 1998. Identification and characterisation of an RPD3 homologue from maize (*Zea mays* L.) that is able to complement an *rpd3* null mutant of *Saccharomyces cerevisiae*. Mol. Gen. Genet. 258, 288–296.

SAS Institute Inc., 1999. SAS/STAT® User's Guide, Version 8. SAS Institute Inc., Cary, NC.

Shachar-Hill, Y., Pfeffer, P.E., Germann, M.W., 1996. Following plant metabolism in vivo and in extracts with heteronuclear two-dimensional nuclear magnetic resonance spectroscopy. Anal. Biochem. 243, 110–118.

Sobolev, A.P., Segre, A., Lamanna, R., 2003. Proton high-field NMR study of tomato juice. Magn. Reson. Chem. 41, 237–245.

Spearman, C.H., 1904. General intelligence objectively determined and measured. Am. J. Psychol. 15, 201–293.

Tobias, R.D., 1995. An introduction to partial least squares regression. In: Proc. SAS User Group, Int. SAS Institute Inc., Cary, pp. 1250–1257.

Wang, X., Kruger, U., Lennox, B., 2003. Recursive partial least squares algorithms for monitoring complex industrial processes. Control Eng. Pract. 11, 613–632.

Ward, J.L., Harris, C., Lewis, J., Beale, M.H., 2003. Assessment of ^{1}H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. Phytochemistry 62, 949–957.

Weckwerth, W., 2003. Metabolomics in systems biology. Ann. Rev. Plant Biol. 54, 669–689.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. In: Krishnaiah, P.R. (Ed.), Multivariate Analysis. Academic Press, New York, pp. 391–420.

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