Bioinformatics Advance Access published April 15, 2004

BIOINFORMATICS

Metabolite fingerprinting: detecting biological features by independent component analysis

M. Scholz[1], *S. Gatzek*[1], *A. Sterling*[2], *O. Fiehn*[1] *and J. Selbig*[1]

[1]*Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany*
[2]*Advion BioSciences Ltd., Norwich NR9 3DB UK*

ABSTRACT

Motivation: Metabolite fingerprinting is a technology for providing information from spectra of total compositions of metabolites. Here, spectra acquisitions by microchip-based nanoflow-direct infusion QTOF mass spectrometry, a simple and high throughput technique, is tested for its informative power. As a simple test case we are using *Arabidopsis thaliana* crosses. The question is how metabolite fingerprinting reflects the biological background.

In many applications the classical principal component analysis (PCA) is used for detecting relevant information. Here a modern alternative is introduced — the *independent component analysis (ICA)*. Due to its independence condition, ICA is more suitable for our questions than PCA.

However, ICA has not been developed for a small number of high dimensional samples, therefore a strategy is needed to overcome this limitation.

Results: To apply ICA successfully it is essential first to reduce the high dimension of the data set, by using PCA. The number of principal components determines the quality of ICA significantly, therefore we propose a criterion for estimating the optimal dimension automatically. The kurtosis measure is used to order the extracted components to our interest.

Applied to our *Arabidopsis thaliana* data, ICA detects three relevant factors, two biological and one technical, and clearly outperforms the PCA.

Key Words: Metabolomics, metabonomics, metabolic profiling, functional genomics, nanospray, feature extraction

Contact: scholz@mpimp-golm.mpg.de

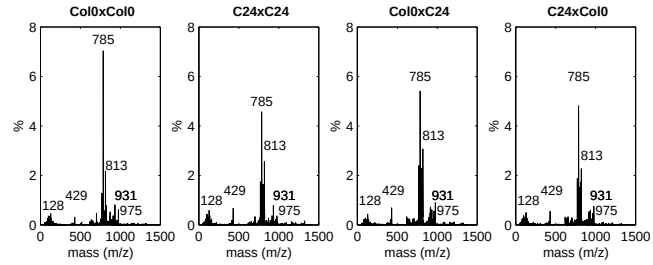

Fig. 1. Mass spectra comparison of different *Arabidopsis* lines and their crosses. The intensities are plotted against the mass (mass-to-charge ratios, *m/z*). From each group one sample is arbitrary taken. The global structure of the spectra is very similar. However there are differences between masses of smaller intensities. To select the relevant information is the challenge for our analysis.

INTRODUCTION

Mutations, natural variation, e.g. ecotypes, and environmental conditions can all influence metabolic processes. Understanding the link between these factors and the overall characteristics of an organism requires a large number of individual analyses. All of these analytical approaches cannot be done by full-composition metabolomic tests, but instead call for a cheaper and faster first-round screening method that groups data according to inherent biological characteristics and distinguishes these from inherent unrelated background noise. Such strategies, without individually determining metabolite identities, have been termed *metabolite fingerprinting* (Fiehn, 2001) and were successfully applied to discriminate strains of bacteria using time-of-flight mass spectrometry (Vaidyanathan *et al.*, 2001) or other techniques such as infrared spectroscopy (Thomas *et al.*, 2000). In biomedical fields, the same strategy is used by applying nuclear magnetic resonance and termed *metabonomics*.

One of the main questions in metabolite fingerprinting is what are the major pieces of information provided by the spectra, and whether the information relates to the experimental conditions or to some interfering signals.

Techniques for visualizing data sets and for extracting important variables in a 'blind' unsupervised way are very helpful for biologists to interpret the given data. Biological background information such as the group affiliations (class labels) are not used in *unsupervised algorithms*. Such techniques are an attempt to present the major or global information given by the data set, unbiased from the experimental target knowledge. An unnoticed supervising effect could appear as well by

adjusting some algorithm-parameters by hand. Therefore, we define different criteria for automatic analysis.

One well-established technique for dimensionality reduction and visualization is the classical *principal component analysis (PCA)*, where the extracted information is represented by a set of new variables, termed *components* or *features*. Diamantaras & Kung (1996) give a good overview of different PCA-algorithms.

In the field of metabolomics PCA became a popular tool for visualizing data sets and for extracting relevant information (Ward *et al.*, 2003; Urbanczyk-Wochniak *et al.*, 2003). However, PCA is only powerful if the biological question is related to the highest variance in the data set. If this is not the case, other techniques of statistics or related fields may be more helpful, depending on the biological question, as shown by Goodacre *et al.* (2003) and Johnson *et al.* (2003) for supervised techniques in combination with validation and pre-processing.

More general questions about the underlying data structure are better investigated by an unsupervised concept of detecting relevant components, unbiased by the experimental target knowledge. Such unsupervised concepts allow a better understanding of how the spectral data reflect the biological experiment. In addition to experimental characteristics, even unexpected factors can be detected. This information can then be used to optimize the experimental conditions.

Different techniques were developed to overcome the disadvantages of the original PCA. Several extensions are in a nonlinear way, such as a nonlinear PCA (Scholz & Vigário, 2002) or locally linear embedding (Roweis & Saul, 2000). However, due to the limited number of samples in high-dimensional data sets, linear alternatives might be more reliable. A very promising linear technique is *independent component analysis (ICA)*. In ICA an independence condition is optimized, which often gives more meaningful components than by optimizing only the variance, as is done by PCA. Because of this the components of ICA are termed *independent components (ICs)*, meaning that different ICs are representing different non-overlapping information.

For applying ICA we assume that the observed data have been determined by some unknown fundamental factors, which are independent from each other. By searching for components as statistically independent as possible these required factors can be detected. These fundamental factors are often termed *sources* and the application field is called *blind source separation, BSS*.

The concept of independent component analysis was first proposed by Comon (1994), with subsequent developments by Bell & Sejnowski (1995). One of the first motivations for ICA was sound signal separation. Currently ICA becomes more important for biomedical applications. Here, applications on time series like EEG data (Makeig *et al.*, 2002) have to be distinguished from applications on rather static data like gene expression (Liebermeister, 2002; Martoglio *et al.*, 2002). There exists a large variety of methods for performing ICA. For time series ICA algorithms such as TDSEP (Ziehe & Müller, 1998) have been developed, whereas algorithms such as FastICA (Hyvärinen & Oja, 2000) are more suitable for static data. Detailed descriptions about ICA are given in Hyvärinen *et al.* (2001) and in Cichocki & Amari (2002).

Usually ICA is applied to data sets having a large number of samples and only a small number of variables. In contrast to that, in metabolite fingerprinting we measure a large number of variables (masses) compared to a relatively small number of samples. Applying ICA directly to this high-dimensional data set is questionable and the results are usually of no practical relevance. One way to avoid this is to reduce the dimensionality before applying ICA. For this, PCA is a suitable technique. We will show that ICA gives optimal results only in connection with PCA as a pre-processing step.

ICA is able to extract as many ICs as the data set has dimensions (number of variables). For technical reasons the ICs have to be sorted, as will be detailed below. In the interesting work of Liebermeister (2002), the ICs were sorted by a combination of a contrast function and a

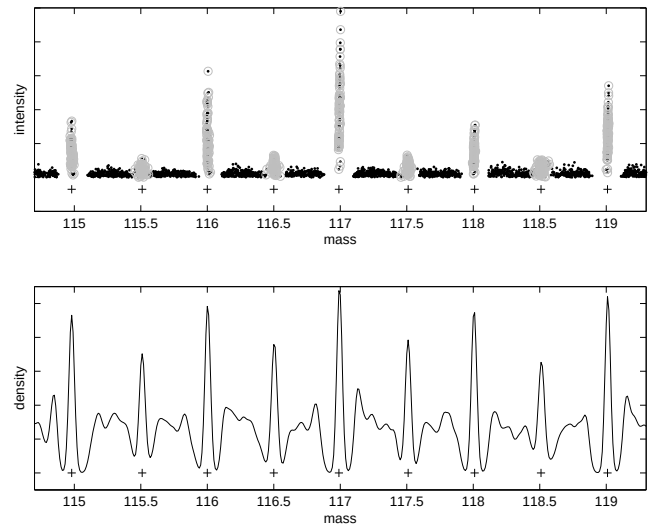

Fig. 2. Combined spectral data. Above, the intensities are plotted against the mass (*m/z*) for all mass-intensity pairs (given by the highest peak in the spectra) over all samples. Only the mass range of 115-119 amu of the total range of 50-1500 amu is shown. For assigning the mass values to a set of variables a density function is used, shown below. The peaks of the density function (marked by a plus '$+$') point to high concentrations of mass values. The masses around one peak (marked by a circle '$\circ$') are assigned to one variable, the residual mass-intensity pairs are removed.

variance criterion. Here we propose to capture the relevant variances by PCA at first and subsequently in ICA the ICs can be extracted and sorted without considering the variance. Here the kurtosis distribution measure is used for sorting.

We set out to test the power of ICA compared to PCA using an easy example: the classification of four genotypes of the model plant *Arabidopsis thaliana*, two parent lines Col0 and C24 and two reciprocal progeny lines with different patterns of maternal inheritance, Col0 x C24 and C24 x Col0. These crosses, the so-called F1 generation, display interesting features such as higher growth, better fitness and improved resistance against biotic and abiotic stress factors. This phenomenon is called *hybrid vigour* or *heterosis* and exploited by classical breeding practices. Therefore, we expected to find the largest distance between the F1 groups and the parents, the second largest difference between the two parents and just a small difference or none at all between the two F1 genotypes.
In a previous study (Taylor *et al.*, 2002), we employed *gas chromatography - mass spectrometry (GC/MS)* with supervised methods to discriminate these crosses. However, when screening large populations of genotypes, GC/MS analysis might be too slow and too expensive, if faster and cheaper methods like fingerprinting would give the same results. We therefore used the same crosses as in the previous work, but now we applied *direct infusion-QTOF mass spectrometric fingerprinting*, as shown in Figure 1.

DATA ACQUISITION

We have 12 samples from each of the four lines or crosses, hence a total of 96 samples. The samples were presented to a direct infusion mass spectrometer without chromatographic separation. An automated nanoelectrospray system was used, that selects the samples automatically from a microtitre plate. Each spectrum reflects the composition of all metabolites, and hence the overall characteristics of one sample, see Figure 1. The highest peaks of the spectra are given as mass-intensity pairs, which have to be assigned to a set of variables, see Figure 2.

Sample preparation. Plant leaf discs of 150 mg FW, homogenized in the frozen state, were extracted with 500 μL acetone/isopropanol 1:1 at -15°C. Cell debris was removed by centrifugation at 14,000 min^{-1}. To 20 μL of the extracts, 20 mL of 70% methanol (acidified with 1% v/v formic acid), and 160 mL of acetone/isopropanol 1:1 were added.

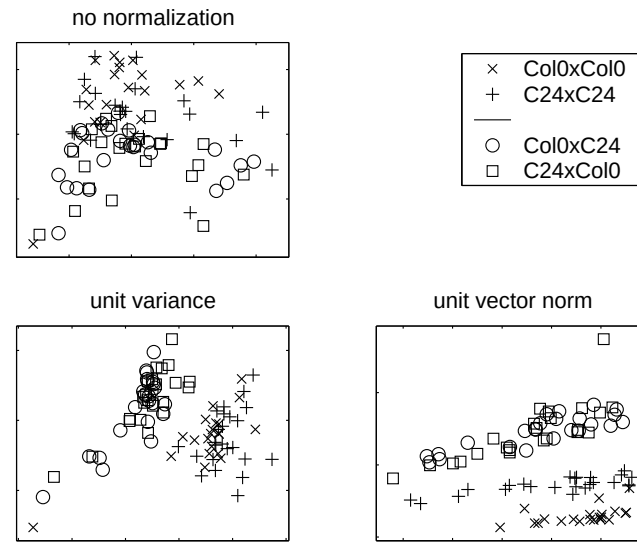

Fig. 3. PCA on normalized data. In each plot the first two components of PCA are plotted against each other. PCA is applied to different normalized data sets. Without any normalization there is no clear separation between the different groups. By scaling the metabolites to unit variance, the parent generation can be separated from the F1 generation. By scaling the samples to unit vector norm, even the parent-lines can be separated.

Automated nanoelectrospray. The NanoMate (Advion BioSciences, Inc. Ithaca, NY), a liquid handling robot, and the ESI Chip (Advion BioSciences, Inc. Ithaca, NY), a microchip consisting of a 10 x 10 array of nanoelectrospray nozzles, together form the automated nanoelectrospray system. The ESI Chip is manufactured from a monolithic silicon wafer by deep reactive ion etching, and other microfabrication techniques (Schultz *et al.*, 2000). Channels extend from the nozzles (8 μm id by 30 μm od) through the microchip to an inlet on the opposing face. Samples are presented to the NanoMate in 96-well microtitre plates. Using proprietary software, ChipSoft (Advion BioSciences, Inc. Ithaca, NY), the NanoMate aspirates sample from the microtitre plate with a disposable conductive pipette tip. The sample is delivered to the next unused inlet with the pipette tip forming a pressure seal around the channel. Nanoelectrospray was initiated by the application of voltage and nitrogen head pressure, and a contact closure was sent, activating the mass spectrometer. At the conclusion of the experiment sample was either returned to the originating microtitre well and the pipette tip ejected, or sample and pipette tip were ejected as one. ChipSoft then selects the next pipette tip and aspirates the requisite volume, repeating the analysis, until the completion of the sample list.

Spectral analysis. NanoMate electrospray/QTOF mass spectra were acquired under the following conditions: 0.09 psi nebulizer gas pressure, 1.43 kV capillary voltage, cone temperature 60°C, 3 μL air gap, 1 s aspiration delay, 2 s voltage delay, 3 s equalization delay, and 2 min mass spectra acquisition from 50 - 1500 amu with a scan rate of 1 amu/ms and a mass resolution of R=6000 FWHM.
A set of mass-intensity pairs is given by the highest peaks in the spectrum from each sample. The peak positions are not identical for each sample. Thus, by combining the mass-intensity pairs of different samples, we have to unify the mass values. The n most significant masses have to be determined, then the mass values of each mass-intensity pair can be assigned to the nearest significant mass.
The significant masses are determined by a density function. High density is related to a significant mass value. The data are also weighted by the intensity. This weighted density function is shown in Figure 2. The function is a Gaussian function with $\sigma = 0.02$ and a sampling rate of 0.01. A number of $n = 1000$ mass values of highest density (highest peaks of the density function) are selected. All data within an interval of $\pm 0.02\ amu$ are labelled by the density-peak-number (1,...,n). These 1000 masses will be referred to as variables.
After removing variables containing missing values, 763 of 1000 variables were left. Due to some unusable spectral-measurements the number of samples was reduced to 92. The complete data set consists of 92 samples separated into 4 groups (two parent-lines and two F1 cross-hybrids). Furthermore, the different masses are represented by 763 variables. The values are mass intensities (*m/z*) given by metabolite concentrations.

NORMALIZATION

In this study the data set was normalized by scaling each sample-vector $x = (x_1, x_2, ..., x_d)$ to unit vector norm $x_{normed} = \frac{x}{||x||}$, d is the number of variables (masses). As vector norm,

$$||x||_p = \sqrt[p]{\sum_i |x_i|^p}$$

the l_2 vector norm ($p = 2$) was taken, but similar results were achieved by using simply the sum of absolute values, termed l_1 vector norm $||x||_1 = \sum_i |x_i|$. The l_2 vector norm is also known as the *Euclidean norm* and can be geometrically interpreted as a projection of the samples x to a hypersphere. The length of this sample vector is scaled to one. By doing this, the ratios between masses do not change. As in nearly each normalization process some information is removed. It is assumed that this information is not relevant to our problem and by removing it, the relevant information can be uncovered. In vector normalization we remove the intensity or power of the sample, which is termed the norm of this sample vector. Because of this, vector normalization is closely related to correlation analysis. Highly correlated samples will be projected close to each other.
Compared to a normalization of the variables (masses) like unit variance, the sample-vector normalization (l_2 vector norm) can achieve better results, see Figure 3. Note the main difference, one is a row-wise and the other is a column-wise normalization. Unit variance will unify the influence of each variable and vector norm will unify the influence of each sample.
PCA and ICA are linear techniques, but the l_2 vector norm is projecting the samples to a curved hypersphere. To avoid this theoretical discrepancy the normalized data set could be linearized by transforming from Cartesian to spherical coordinates, where the angles have a kind of hierarchical order. A symmetric transformation might be more suitable, which can be done by using the arcsine element-wise $x_{linear} = \arcsin(x_{l_2 normed})$. However, the influence on the ICA result is relatively small, and to simplify matters, it is not further considered in this article.

PCA – PRE-PROCESSING

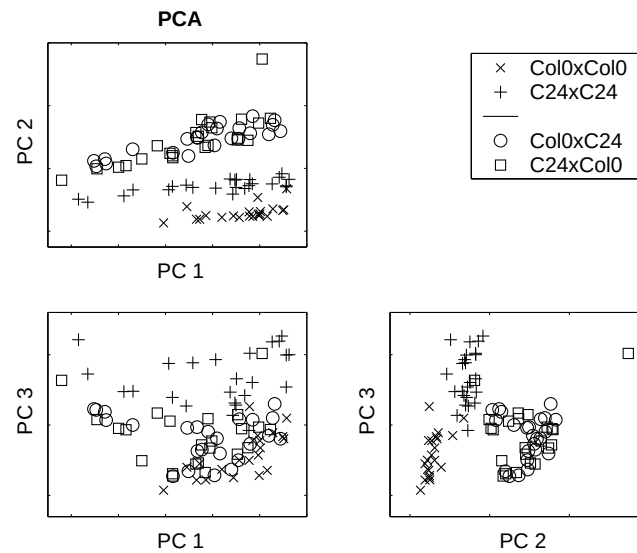

Fig. 4. PCA on vector normalized data. The first three principal components (PCs) are plotted pairwise against each other. Note that the first PC (of highest variance) is not related to our problem of separating the sample groups. Better results are given by components of smaller variance, PC 2 and PC 3.

The criterion that is optimized in PCA is the variance. By applying PCA for visualization we have to assume that the most interesting information is directly related to the highest variance in the data. The best projection or visual-

ization is then given by the first two principal components (PCs) of highest variance. Often this assumption does not hold and we find interesting projections in later components, see Figure 4. We can still assume that the interesting information is related to a significantly high amount of variance but not to the highest amount. PCA can still be used to reduce the high dimension of the data to a relatively low dimension, which covers all relevant variances. Such a pre-processing step should preserve all of the relevant information and reduce only the noise given by small variances.

ICA – INDEPENDENT COMPONENT ANALYSIS

Once the relevant variances are discovered by PCA, on this lower dimensional data set, a technique can be applied which optimizes other criteria than the variance. A promising technique for this is the independent component analysis - ICA.

Similarly to PCA, ICA gives a set of components. In contrast to PCA these components are constructed such as to minimise the dependence and are therefore termed *independent components (ICs)*. Independence is a stronger condition than uncorrelation in PCA. This allows detection of more meaningful components than by PCA. These components are not restricted to be orthogonal in ICA.

To achieve independent components, different criteria (contrast functions) can be optimized: higher-order dependencies, entropy or kurtosis.

In this article, ICA was performed by the widely-used FastICA - algorithm (Hyvärinen & Oja, 2000).

SIGNIFICANT COMPONENTS - KURTOSIS

ICA is able to extract as many components as the data set has dimensions. These components have no order. For practical reasons we have to define a criterion for sorting these components to our interest. One measurement which can match our interest very well, is the kurtosis.

Kurtosis is a classical measure of non-Gaussianicity, and is computationally and theoretically relative simple. It indicates whether the data are peaked or flat relative to a Gaussian (normal) distribution. A Gaussian distribution has a kurtosis of zero. Positive kurtosis indicates a 'peaked' distribution (super-Gaussian) and negative kurtosis indicates a 'flat' distribution (sub-Gaussian).

$$kurtosis(z) = \frac{\sum_{i=1}^{n}(z_i - \mu)^4}{(n-1)\sigma^4} - 3$$

where $z = (z_1, z_2, ..., z_n)$ is representing a variable or component with mean μ and and standard deviation σ, n is the number of samples. The kurtosis is the fourth auto-cumulant after mean (first), variance (second), and skewness (third).

From purely Gaussian distributed data no unique independent components can be extracted, therefore, ICA should only be applied to data sets where we can find components that have a non-Gaussian distribution.

Examples of super-Gaussian distributions (highly positive kurtosis) are speech signals, because these are predominantly close to zero. However, for metabolite data sub-Gaussian distributions (negative kurtosis) are more interesting. Negative kurtosis can indicate a cluster structure or at least an uniformly distributed factor. The former can resolve between two experimental conditions (high and low concentrations of metabolites), whereas the latter can represent a continuously changed experimental factor such as the temperature or the light intensity. Thus the components with most negative kurtosis could give us the most relevant information.

ICA ON EXPERIMENTAL DATA

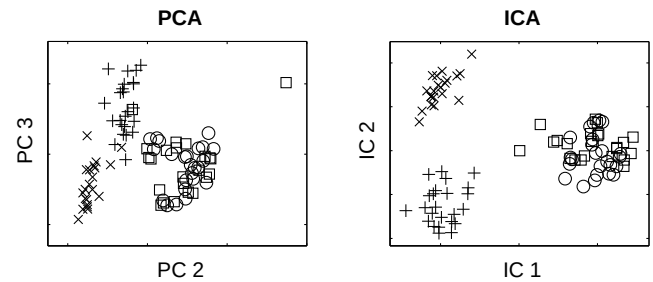

Fig. 5. ICA compared to PCA. The best PCA-visualization given by PC 2 and PC 3 is plotted on the left. The different groups are only partially separable. Compared to this the ICA result, given by the two independent components of most negative kurtosis, IC 1 and IC 2, is shown on the right. ICA gives a projection of the data with a greater separation between the different groups.

First, the data set is normalized to unit vector norm (l_2 vector norm). Second, PCA is used as a pre-processing step for reducing the dimensionality to 6 dimensions, see next section for details. The FastICA algorithm is applied and the extracted independent components are sorted by their kurtosis.

Figure 5 shows the results of the ICA compared to PCA. The two components of ICA with the most negative kurtosis, IC 1 and IC 2, clearly gave a greater separation of the sample groups. Furthermore, the two independent components have a biologically independent interpretation. The first component separates between the parent and the F1 generation. The second component separates between the parent lines.

As shown in Figure 7, three components with clearly negative kurtosis are detected. The third component (IC 3), an

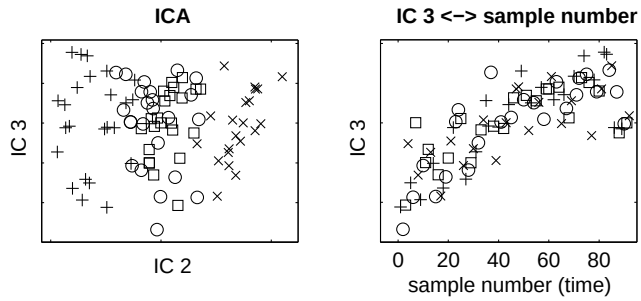

Fig. 6. The third component of ICA (ic 3) has no information about the experimental groups (left). However, there is a relation to the time, when the samples are measured, shown on the right. This technical factor could not be detected by PCA.

almost uniformly distributed factor, has no relation to the groups of our biological experiment. However, we could detect a relation to the identifier of the samples, representing the order over time measured in the mass spectrometer, see Figure 6. Hence IC 3 is an unintended technical factor. The reason could be chemical contamination over time.

OPTIMAL PCA-DIMENSION

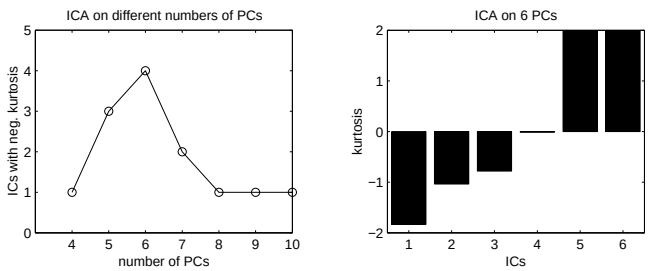

Fig. 7. Left: Different numbers of principal components are used for dimensionality reduction. ICA is applied for each of these reduced data sets. Plotted are the number of extracted ICs with negative kurtosis. By using the first 6 components of PCA, ICA can extract the highest number of interesting ICs, whereas the kurtosis of IC 4 is close to zero. Right: For this 6 dimensional reduced data set, the kurtosis of all extracted ICs are plotted.

By using the PCA as a pre-processing step, the number of PCs, hence the optimal reduced dimensionality is usually unknown. Thus we have to find a way to estimate this dimension. Here, the estimation is aligned with the aim of our analysis, i.e. to find as many relevant components as possible. As a negative kurtosis indicates relevant components, the dimension, where we can extract the highest number of independent components with negative kurtosis is the optimal dimension.
On our experimental data set we found optimal results with 5, 6, and 7 PCs. Although ICA detects even 4 relevant components using 6 PCs, see Figure 7, the fourth is close to zero and so should not be counted. The first two ICs are plotted in Figure 5. If the number of PCs is too small (< 5), relevant information will be removed by PCA. If the number of PCs is too high (> 7) the higher level of noise masks the relevant information. In both cases the component for separating the parents will be lost.
Alternative to counting simply the number of components with negative kurtosis, the square sum over these negative values can be used. This might be a more reliable criterion, as a kurtosis close to zero has little effect. From the point of information theory a measure based on entropy could be used as well. However, such criteria are needed for an automated analytical procedure.

INFLUENCE VALUES

PC 2 P ↔ F1		**PC 3** $P_{Col} \leftrightarrow P_{C24}$		**IC 1** P ↔ F1		**IC 2** $P_{Col} \leftrightarrow P_{C24}$	
mass	**infl.**	**mass**	**infl.**	**mass**	**infl.**	**mass**	**infl.**
769.2	0.46	785.1	-0.38	913.3	0.44	769.2	0.52
913.3	0.28	819.2	-0.26	911.2	0.29	803.2	0.42
915.2	0.25	769.2	0.24	915.2	0.24	770.2	0.25
770.2	0.21	913.3	-0.22	914.3	0.24	804.2	0.25
975.1	0.20	803.2	0.21	931.1	-0.20	975.1	-0.19
959.2	0.20	786.1	-0.17	912.2	0.18	819.2	-0.15
785.1	-0.19	820.2	-0.16	794.2	0.17	913.3	-0.15
797.2	0.19	911.2	-0.16	889.3	0.15	797.2	0.13
911.2	0.18	88.0	0.15	932.1	-0.14	911.2	-0.12
914.3	0.16	108.0	0.15	778.2	0.12	771.2	0.12

Table 1. Mass influence. The 10 masses of highest influence are shown for different components. On the left the masses given by the classical PCA are shown for PC 2 and PC 3. These are the PCs which are closest to the first two ICs of ICA, shown on the right. The masses given by ICA are different to these of PCA and are rather assignable to only one IC. These higher mass separations are shown in Figure 8.

If the components represent biological factors, it will be important to know which masses are highly involved. Thus the influences of each mass on each of the components have to be calculated, which is often done in a similar way for PCA. Such influence values are often termed *loadings* or *weights*.
We would expect that most of the masses are related to specific biological factors, and hence the masses should be assignable to different components. Only a small set of masses (meta-masses) should be caused by several biological activities, having a significant influence on more than one component.

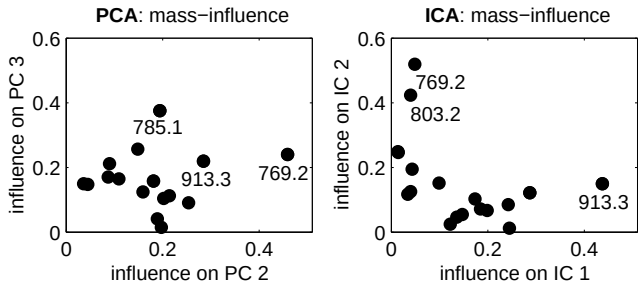

Fig. 8. Mass influences. For each mass from Table 1 the absolute influence on each component is plotted. The masses in PCA have a greater influence on both components than the masses in ICA, which are assigned more to one or to the other component.

In Table 1 the influences are given for components of ICA and PCA, which are not identical. In Figure 8 it becomes apparent that ICA fits the expectation better, as different masses are linked to different independent components. This can indicate that components of ICA give us a better description of the biological factors.

The influence values are given by the transformation matrices of PCA and ICA. PCA transforms a d-dimensional sample vector $x = (x_1, x_2, \ldots, x_d)^T$ into a usually lower dimensional vector $y = (y_1, y_2, \ldots, y_k)^T$, where d is the number of masses and k is the number of selected components. The PCA transformation is given by the eigenvector matrix V, $y = Vx$. Similarly, ICA transforms this vector y to the required vector $z = (z_1, z_2, \ldots, z_k)^T$, containing the independent values z_i for each IC i. For that a de-mixing matrix W is estimated by ICA, $z = Wy$. V gives the influences of each variable (mass) on each of the PCs, whereas W gives us the influence of each PC on each of the ICs. We can combine both matrices $U = W * V$ to a direct transformation $z = Ux$, where U gives vector-wise the required influences of each variable on each of the ICs. For comparing these influences with the influences of solely applying PCA (Table 1), the length of the influence vector has to be scaled to one (l_2 vector norm).
Note that informative influence values can only be determined if the data set is adequately normalized beforehand. It is also required that the variables have a mean of zero.

ICA does not attempt to cluster variables and thus is not a cluster algorithm. However the variables can be assigned to different components with respect to their influence values, as is done in Table 1. The results are closely related to an overlapping clustering. Based on these influence values, distances between masses could be calculated, and used for network analysis.

OUTLIER DETECTION

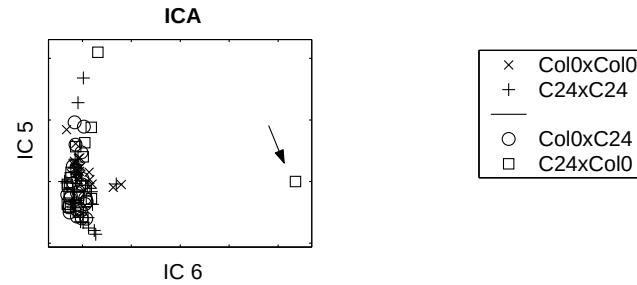

Fig. 9. Outlier detection by ICA. The last two components with the most positive kurtosis are plotted against each other. The IC 6 clearly indicates an outlier, marked by an arrow.

The sensitivity of ICA to outliers can be seen as an advantage. The outliers are indicated by a component with high positive kurtosis. Thus ICA can be used to remove outlier-samples or to correct those by moving it to the residual samples in the direction of this component.
The assignment of an outlier to a separate component (IC 6) reduces the negative effect to the required components (IC 1 and IC 2). Thus, even without removing outliers, ICA gives good results.

CONCLUSION

The widely used assumption that the desired information is related to the highest variance in the data set does not hold for our experiment. Thus sufficient results can not be obtained by solely optimizing the variance, as is done by PCA.
We have shown that ICA has a higher informative power when it is combined with suitable pre-processing and evaluation criteria. More precise biological features are detected, and even a technical factor was found to influence the spectral information, which could not be detected by PCA.
The kurtosis measure clearly denotes three independent components as significant. The first is usable for separating the *Arabidopsis* crosses from the background parental lines, and the second contains information for discriminating the two parental lines. The third component could be interpreted as a contribution of chemical noise due to increasing contamination of the QTOF skimmer along the analytical sequence.
ICA, together with the proposed criteria, forms an automated analytical procedure that offers a metabolite fingerprinting technique designed for high sample throughput. We will make the approach described in this study available to the public by integrating it into *MetaGeneAlyse* (Daub *et al.*, 2003), a web-based tool for analyzing bio-

logical data from metabolomics, proteomics and transcriptomics.

ACKNOWLEDGMENTS

The authors thank Thomas Altmann and Rhonda Meyer for initiating and stimulating the *Arabidopsis hybrid vigour* (heterosis) project, which aims to use recombinant inbred and near isogenic lines for functional genomics. We also thank Wolfram Weckwerth, Grit Rother, Stefanie Hartmann, and John Lunn for valuable discussions.

REFERENCES

Bell, A. J. & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, **7**, 1129–1159.

Cichocki, A. & Amari, S. (2002). Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. Wiley.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, **36**, 287–314.

Daub, C. O., Kloska, S. & Selbig, J. (2003). MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics*, 2332–2333. http://metagenealyse.mpimp-golm.mpg.de/

Diamantaras, K. & Kung, S. (1996). Principal Component Neural Networks. Wiley, New York.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genom.*, 155–168.

Goodacre, R., York, E. V., Heald, J. K. & Scott, I. M. (2003). Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry*, **62**, 859–863.

Hyvärinen, A., Karhunen, J. & Oja, E. (2001). Independent Component Analysis. J. Wiley.

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, **4–5**, 411–430.

Johnson, H. E., Broadhurst, D., Goodacre, R. & Smith, A. R. (2003). Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry*, **62**, 919–928.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 51–60.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E. & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science*, **295**, 690–694. In Reports.

Martoglio, A. M., Miskin, J. W., Smith, S. K. & MacKay, D. J. C. (2002). A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*, **18**, 1617–1624.

Roweis, S. & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323 – 2326.

Scholz, M. & Vigário, R. (2002). Nonlinear PCA: a new hierarchical approach. In Verleysen, M., (ed.) *Proceedings ESANN*. pp. 439–444.

Schultz, G. A., Corso, T. N., Prosser, S. J. & Zhang, S. (2000). A fully integrated monolithic microchip electrospray device for mass spectrometry. *Anal. Chem.*, 4058–4063.

Taylor, J., King, R. D., Altmann, T. & Fiehn, O. (2002). Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics*, **18**, 241–248.

Thomas, N., Goodacre, R., Timmins, É. M., Gaudoin, M. & Fleming, R. (2000). Fourier transform infrared spectroscopy of follicular fluids from large and small antral follicles. *Human Reproduction*, **15**, 1667–1671.

Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L. & Fernie, A. R. (2003). Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO reports*, **4**, 989–993.

Vaidyanathan, S., Rowland, J. J., Kell, D. B. & Goodacre, R. (2001). Discrimination of aerobic endospore-forming bacteria via electrospray-ionisation mass spectrometry of whole cell suspensions. *Anal. Chem.*, **73**, 4134–4144.

Ward, J. L., C. Harris, J. L. & 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.

Ziehe, A. & Müller, K.-R. (1998). TDSEP - an efficient algorithm for blind separation using time structure. In *Proc. ICANN'98, Int. Conf. on Artificial Neural Networks*. pp. 675–680.