Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS

Hongmei Lu, Warwick B. Dunn, Hailin Shen, Douglas B. Kell, Yizeng Liang

Traditional options available for deconvolution of data from gas chromatography-mass spectrometry (GC-MS) experiments have mostly been confined to semi-automated methods, which cannot compete with high-throughput and rapid analysis in metabolomics. In the present study, data sets acquired using GC with time-of-flight MS (GC-TOF-MS) were processed using three different deconvolution software packages (LECO ChromaTOF, AMDIS and SpectralWorks AnalyzerPro).

We paid attention to the extent of detection, identification and agreement of qualitative results, and took interest in the flexibility and the productivity of these programs in their application. We made comparisons using data from the analysis of a test-mixture solution of 36 endogenous metabolites with a wide range of relative concentration ratios.

We detected differences in the number of components identified and the accuracy of deconvolution. Using the AMDIS Search program, the resulting mass spectra after deconvolution were searched against the author-constructed retention index/mass spectral libraries containing both the mass spectra and the retention indices of derivatives of a set of metabolites. We based analyte identifications on both retention indices and spectral similarity.

The results showed that there were large differences in the numbers of components identified and the qualitative results from the three programs. AMDIS and ChromaTOF produced a large number of false positives, while AnalyzerPro produced some false negatives. We found that, in these three software packages, component width is the most important parameter for predicting the accuracy of the deconvoluted result.

Keywords: Component search; Gas chromatography-mass spectrometry; GC-MS; Metabolomics; Software; Spectral deconvolution

Hongmei Lu*, Yizeng Liang
College of Chemistry and
Chemical Engineering,
Central South University,
Changsha 410083, P. R. China

Warwick B. Dunn, Hailin Shen, Douglas B. Kell
Manchester Centre for
Integrative Systems Biology and
Bioanalytical Sciences Group,
Manchester Interdisciplinary
Biocentre,
The University of Manchester,
131 Princess Street,
Manchester M1 7DN, U. K.

*Corresponding author.
Tel.: +86 731 8830824;
Fax: +86 731 8830831;
E-mail:
hongmeilu@mail.csu.edu.cn

1. Introduction

Metabolomics is an emergent scientific discipline which is applied to many different applications. It is becoming a tool in the functional annotation of genes and enzymes and in the comprehensive understanding of cellular and organism-specific responses to biological, environmental and drug-related perturbations. Metabolomics has been defined as the unbiased identification and quantification of all metabolites in a biological system [1–3]. Metabolomics provides a number of advantages when compared to studies of the transciptome and proteome [4–7].

The number of metabolites present in an organism is large. *Saccharomyces cerevisiae* contains approximately 600 metabolites [8], the plant kingdom has an estimated 200,000 primary and secondary metabolites [2] and the human metabolome contains approximately 1500 metabolites [9], excluding lipids and exogenous metabolites derived from food or pharmaceuticals. Moreover, differences in metabolite concentrations are observed to be greater than 5000-fold, although these differences are not proportional to the biological relevance of these metabolites [10]. Global analysis of so many metabolites with divergent physical properties and large dynamic concentration ranges is a great, and currently unresolvable, challenge to analytical techniques, data processing and data interpretation. As an alternative, metabolic profiling is commonly applied to detect a wide array of metabolites, related by chemical properties or metabolite class, in any given sample.

There are a number of analytical tools used to analyze these highly complex mixtures [2–6,11]. These include gas chromatography-mass spectrometry (GC-MS),

0165-9936/$ - see front matter © 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.trac.2007.11.004

liquid chromatography-mass spectrometry (LC-MS), capillary electrophoresis-mass spectrometry (CE-MS), Fourier transform infrared spectroscopy (FT-IR) and nuclear magnetic resonance spectroscopy (NMR). Some advantages of GC-MS include stable retention time, robust protocols for sample preparation and instrument operation and the ability to identify metabolites by assessing retention time/index and electron-impact mass spectra. A relatively broad coverage of compound classes can be detected with good sensitivity, after appropriate instrument optimization [12,13], including organic and amino acids, sugars, sugar alcohols, phosphorylated intermediates and lipophilic compounds. GC-MS can be regarded as the gold standard for metabolic profiling [14].

In GC-MS, sample components are expected to exit the chromatography column and be introduced into the MS one-by-one. However, compounds often travel through the column with similar speeds, so a large number of the compounds coelute or are not completely resolved chromatographically. To obtain accurate pure-mass spectra of a specific compound in coeluted peaks in order to identify and quantify compounds correctly in metabolomics, mathematical multivariate curve-resolution procedures, (often named deconvolution) need to be applied. Multivariate curve-resolution methods that have been developed include iterative [15,16], non-iterative [17,18], and hybrid approaches [19,20]. They have been used to clarify chromatographic and spectral profiles from overlapping chromatographic peaks obtained using various types of hyphenated chromatography systems [21]. The main disadvantage of the methods developed is that they are very difficult to automate for different complex biosamples that have specific analytical needs.

Metabolomics generates floods of data every day [22]. It is clear that deconvoluting metabolomics data with conventional manual methods is too time-consuming and tedious, and requires skilled individuals. However, the increasing capability of chromatography-MS systems, particularly improved signal-to-noise (S/N) ratios and higher peak capacity, means that the analyst expects to be able to analyze in a single analysis hundreds of metabolites whose chemical nature is unknown (60–90% of the total in a complex matrix) [23]. The extremely complex samples inevitably lead to differences in peak shapes, retention-time drift, and variation in the response for different compounds, which make deconvolution more complex and difficult.

As a result, deconvolution is a major bottleneck of metabolomics. The development of metabolomics requires specialized, automated software or tools for deconvolution after high-throughput instrumental analysis. In recent years, tools have been developed to address the problems of co-eluting interferences, and to identify accurately as many peaks as possible. Instrument manufacturers (e.g., LECO, Waters, Shimadzu and Agilent Technologies) and third-party suppliers of data-analysis software (e.g., AMDIS, AnalyzerPro and XCMS) have provided deconvolution functions in software packages.

We wondered whether these software packages are capable of deconvoluting metabolomics data and whether their results are credible and reliable. To our knowledge, there has been no broad comparison of these software packages. For this article, to evaluate and compare software packages in an applied situation, we prepared a standard data set with a specifically chosen standard mixture in known proportions. The data were processed using three separate programs – AMDIS (NIST), ChromaTOF (LECO) and AnalyzerPro (Spectralworks) – with GC-MS data from a GC-TOF-MS instrument (LECO). ChromaTOF software was obtained as part of the LECO Pegasus III TOF-MS instrument that we bought. AMDIS was downloaded free of charge. AnalyzerPro was a one-month free demo from Spectralworks Ltd, UK. No other software packages were used because they were unavailable or our knowledge was limited. In view of our on-going metabolomics studies, we investigated these three data-deconvolution-software packages to increase our options for data analysis.

2. Materials and methods

2.1. Preparation of analytical metabolite solutions

Some 36 single-metabolite solutions with an approximate concentration of 20 mM were prepared in 50:50 water:acetonitrile (Metabolite Stock Solutions 1–36) (as described in Table 1). The metabolites are typical endogenous components and include common metabolites (e.g., amino acids, organic acids, sugars, sugar alcohols and aromatic amines). Combinations of these metabolite solutions were prepared for analysis. Metabolite Stock Solutions 1–36 were diluted and mixed to produce Solutions 1–4, with concentrations of 500 μM, 350 μM, 150 μM and 50 μM, respectively, containing all 36 metabolites. Solution 5 was prepared with variations in metabolite concentrations, 50% of the metabolites were at a concentration 10 times greater than the other 50% (i.e. 50 μM and 500 μM, respectively). Solutions 1–5 were lyophilised (HETO VR MAXI vacuum centrifuge attached to a HETO CT/DW 60E cooling trap; Thermo Life Sciences, Basingstoke, UK) before chemical derivatization and analysis.

2.2. GC-TOF-MS

Two-stage chemical derivatization was performed prior to GC-TOF-MS analysis. First, oximation was performed by heating the samples with O-methylhydroxylamine (50 μL; 20 mg/ml in pyridine; 80 min; 40°C); then, the samples were trimethylsilylated with MSTFA (N-acetyl-N-(trimethylsilyl)-trifluoroacetamide; 50 μL; 80 min; 40°C).

Table 1. List of metabolites

Pyruvic acid	Valine
Leucine	Sorbic acid
Proline	Threonine
Fumaric acid	Nicotinic acid
Uracil	4-hydroxyproline
Aspartic acid	2-hydroxyglutaric acid
Oxaloacetic acid	Arabinose
Ribitol	Rhamnose
2-oxoglutaric acid	Asparagine
Fructose	Trans-aconitic acid
Glucose	Citric acid
Glucuronic acid	Gluconic acid
Quinaldic acid	N-acetylglucosamine
Glucose-6-phosphate	Indole-3-acetic acid
Serotonin	5-hydroxytryptophan
Glutaric acid	Citramalic acid
Myo-inositol	Ascorbic acid
Tryptamine	Sucrose

Derivatized samples were analyzed by GC-TOF-MS using a Agilent 6890 N GC instrument (Stockport, UK) coupled to a LECO Pegasus III MS instrument (St Joseph, USA), controlled with ChromaTOF software v2.15. Sample analysis was randomized and three machine replicates were performed for each sample.

The GC instrument was operated in split mode using helium as carrier gas in constant-flow mode, with an initial GC temperature of 70°C. A DB-50 GC column (Supelco, Gillingham, UK; 30 m × 0.25 mm × 0.25 μm film thickness) was used. The transfer-line and source temperatures were 250°C and 230°C, respectively. The mass range used was 30–600 Da with a detector voltage of 1700V. Each sample was analyzed using two sets of instrument conditions, A and B – A was as previously described [12] and B was identical to A with the exception of the oven-temperature program, which was increased from 28°C/min in A to 70°C/min in B. All data were exported as netCDF files for further data analysis. The A data were mainly to be used for manual deconvolution to establish the judgment rule, while the B data were to be used for evaluating the software packages.

2.3. Data processing

In this study, we used an Acer computer with two Pentium (R) D 3.0 GHz processors and 2 GB RAM for Windows-based applications. The deconvolution-software packages were operated with Windows XP Professional (Version 2002, Service pack 2).

2.3.1. AMDIS. The Automated Mass Spectral Deconvolution and Identification System (AMDIS, Version 2.64, NIST, US) extracts spectra for individual components in a GC-MS data file and identifies compounds by matching these spectra against specialized libraries or the NIST 02 library. It was developed at NIST with support from the US Department of Defense and is freely available.

We subjected GC-MS data files to analysis by AMDIS in simple mode. Data deconvolution was performed with the following specifications:

- component width = 12;
- adjacent peak subtraction = 1;
- resolution = medium;
- sensitivity = medium; and,
- shape requirements = medium.

2.3.2. ChromaTOF. ChromaTOF software (version 2.15) was available as part of the LECO Pegasus III TOF-MS instrument. In the ChromaTOF software, the settings of parameters derived from our previous study and experience [12]:

- the S/N threshold was set at 10;
- baseline offset at 1.0;
- data points for averaging at 3; and,
- peak width at 2.5.

2.3.3. AnalyzerPro. AnalyzerPro demo (Version 2.0.0.1) is a vendor-independent software, which is provided by Spectralworks Ltd, UK. AnalyzerPro can analyze a GC-MS file using qualitative processing to detect components using proprietary algorithms. The spectra for the components found are automatically enhanced, eliminating the need for manual background subtraction or further spectral refinement algorithms. AnalyzerPro can import a number of vendors' file formats (e.g., ABI/MDS Sciex, Agilent, JCamp, NetCDF, MassLab, and Thermo Electron) and convert them to .swx files that are optimized for data analysis. We performed deconvolution with the following specifications suggested by the software company:

- minimum masses = 6;
- area threshold = 500;
- height threshold = 200;
- width threshold = 0.02 min;
- resolution = low;
- scan window = 2;
- S/N = 5; and,
- smoothing = 1.

2.3.4. NIST MS search software. NIST MS Spectral Search Program (version 2.0 a), distributed by the Standard Reference Data Program of NIST, was used to compare software-deconvoluted MS result files with the standard mass spectra in our own reference libraries, University of Manchester (UoM) Yeast GC-TOF-MS Library containing both retention indexes of derivatives of a set of metabolites (as determined under our conditions) and the corresponding mass spectra. UoM Yeast GC-TOF-MS Library includes 254 mass spectra of the oxime-TMS derivatives of metabolites. The quality of data deconvolution is described by the number of the components detected and the accuracy of the deconvoluted mass spectrum. The match factor is a weighted count describing how well the deconvoluted spectrum matches the theoretical spectrum of a metabolite candidate in the library. Library searching uses the normal identity-search mode. The match factor of the full mass spectrum for the deconvoluted components with the standard mass spectrum in reference libraries was taken as the first identification criterion that provided an indication of the reliability of assignment. The setting of the match-factor threshold was based on the statistical results of the standard mass spectra in our own reference libraries and the deconvoluted spectra. When the match factor was 850, 90% of the qualitative results of the deconvoluted spectra from manual methods were accurate. This value should be considered acceptable.

In the following study, we set the threshold of the match factor at 850. The greater the number of mass spectra with match factor greater than 850, the closer the deconvoluted result is to the true result.

The retention index was taken as a second criterion. To determine the Kovats index, we used *n*-alkanes (*n*-dodecane, *n*-pentadecane, *n*-nonadecane and *n*-docosane).

3. Results

3.1. Samples with different combinations of concentration

It has been noted that the method of derivatization can produce more than one derivative for a single metabolite [24,25], so the number of components detected does not equal the number of metabolites before derivatization. As we had no criteria on how many components exist in solution after derivatization, in an attempt to obtain a ''true'' measure of the number of metabolite derivatives in the chromatogram, we integrated the judgment of an experienced GC-MS analyst with the result from three data sets of Solution 1 analyzed with a slow temperature ramp (28°C) with our manual methods HELP (Heuristic Evolving Latent Projections) and SCC (Spectral Correlative Chromatography) [17,18,26,27], which had proved to be very useful for analyzing and comparing complex hyphenated chromatography data, to obtain a relatively reliable number of components in samples for the following evaluation of the software packages. With those methods, 51 metabolite derivatives were validated in solutions and were used as the standard for the following evaluation work.

We evaluated the comparative performance of the software packages in the analytical data with two metabolite sample solutions covering a range of concentrations (see Section 2.1). The deconvolution parameter settings are given in Section 2.3. An example of the deconvoluted results from the three software packages, employing the retention time window 378–388s of Solution 1, is shown in Fig. 1a. The concentration of metabolites is 500 μM. This retention-time window included four metabolite derivatives (uracil, nicotinic acid, glutaric acid and citramalic acid) and 1 *n*-alkane (*n*-pentadecane). The deconvoluted results showed that many extraneous, aberrant components from system noise at the same fragment were automatically deconvoluted, although 4 standards and 1 *n*-alkane are deconvoluted by all the software packages, as expected.

The kind of error that the data encountered is particularly evident in the result from AMDIS. One single peak was deconvoluted as multiple components. AMDIS deconvoluted 48 components out of the 4 metabolite derivatives (uracil, nicotinic acid, glutaric acid and citramalic acid) and 1 *n*-alkane (*n*-pentadecane). By comparison, AnalyzerPro produced the least false positives (i.e. only deconvoluted 8 components out of 4 metabolite derivatives) and 1 *n*-alkane with correct deconvoluted mass spectra (Fig. 1a).

The complete data set was then analyzed using the three software packages and the deconvoluted and match results acquired are listed in Table 2.

From the results in Table 2, we found that:

- all metabolite derivatives in solutions with different concentration were detected by AMDIS;
- 8 metabolite derivatives were not detected by ChromaTOF when the concentration decreased to 50 μM; and,
- 2–38 metabolite derivatives were not detected by AnalyzerPro in solutions with four different concentrations.

However, the number of the deconvoluted components for sample solutions with different concentrations using AMDIS reached 522–720 (i.e. AMDIS deconvoluted several hundred components, including the 51 true components). The number is far greater than using ChromaTOF (78–173) and AnalyzerPro (14–67). This demonstrated that AMDIS produced more false positives than the other two software packages. It introduced another tough question that we have no way to answer – ''How does the analyst pick out the correct deconvoluted spectra from those results without previous knowledge, even though the spectra of metabolites are detected and deconvoluted correctly?''

AnalyzerPro and ChromaTOF provided the least false positives and therefore made it easier to define true metabolites, as the numbers of components detected by those two software packages were closer to the true number of metabolite derivatives in solution. However, some metabolites were not detected, and that meant they produced false negatives. So far, none of these three software packages has provided a good balance between avoiding false positives and avoiding false negatives. If we provisionally ignore false positives and

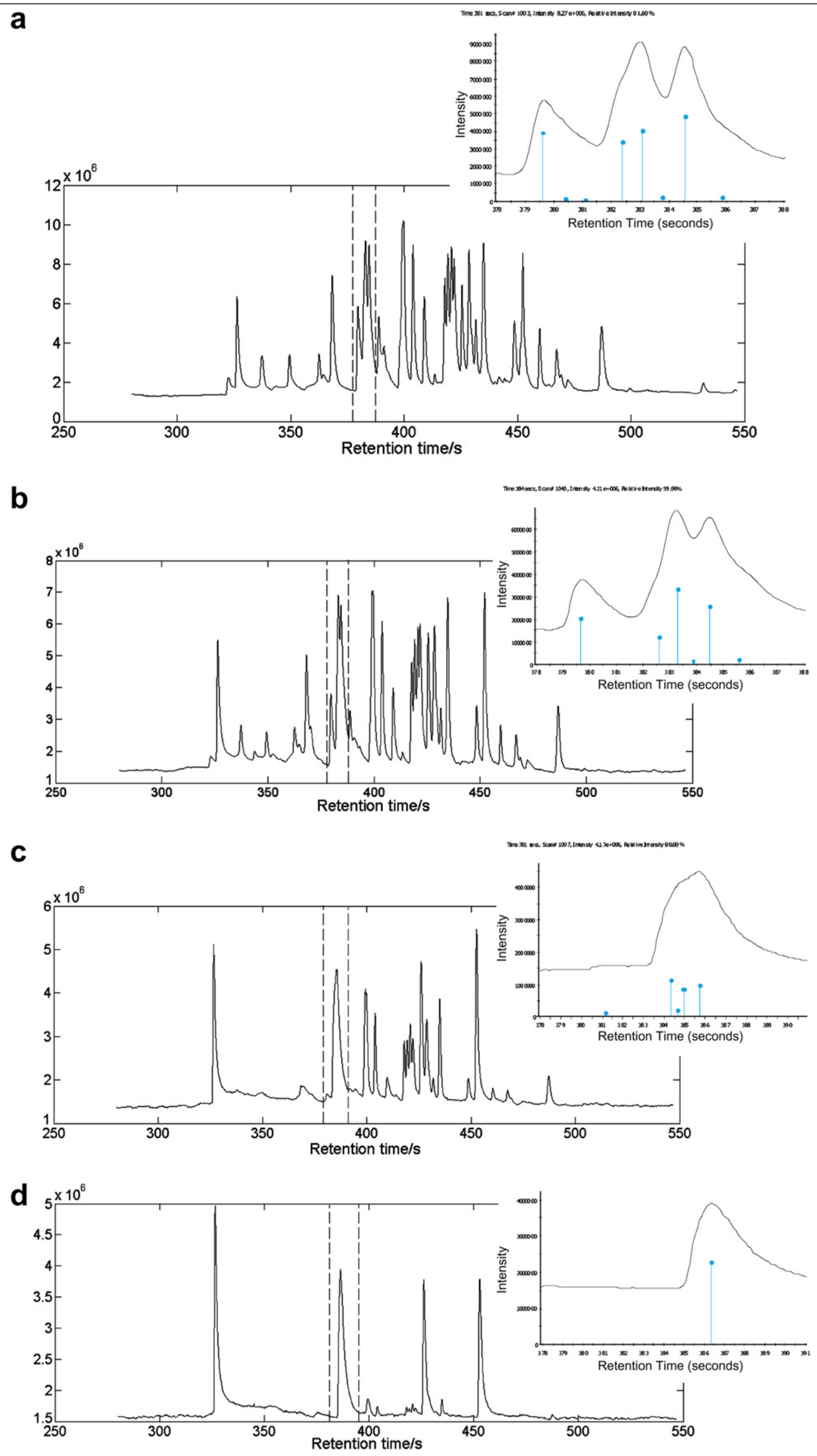

Figure 1. Zoom in on the retention-time window of 378–388 s and the locations of deconvoluted components from AnalyzerPro, and TIC of: (a) Solution 1; (b) Solution 2; (c) Solution 3; and, (d) Solution 4.

Table 2. Results acquired with 3 software packages from samples with different combinations of concentration

Data		Solution 1	Solution 2	Solution 3	Solution 4	Solution 5
Number of components deconvoluted	ChromaTOF	173	161	121	78	162
	AMDIS	720	620	529	522	720
	AnalyzerPro	67	49	38	14	42
Number of metabolite derivatives undetected	ChromaTOF	0	0	0	8	0
	AMDIS	0	0	0	0	0
	AnalyzerPro	2	9	17	38	19
Number of metabolite- derivative spectra deconvoluted correctly	ChromaTOF	37	31	28	14	27
	AMDIS	32	30	20	8	26
	AnalyzerPro	28	24	14	5	18

false negatives produced by the software packages, the validity of ChromaTOF for deconvoluting spectra was best. The number of metabolite spectra deconvoluted correctly by ChromaTOF was greater than the numbers deconvoluted correctly by AMDIS and AnalyzerPro (Table 2).

In addition, with the concentration of the components decreasing from 500 μM to 50 μM, we found that the number of deconvoluted components decreased, the number of undetected compounds increased and the number of metabolite spectra deconvoluted correctly decreased. Taking the counterparts of the retention-time window of 378–388 s in Solution 1 from sample solutions of different concentrations to show the effect of the concentration, these fragments still include 4 metabolites (uracil, nicotinic acid, glutaric acid and citramalic acid) and 1 *n*-alkane (*n*-pentadecane). Only the result for AnalyzerPro is listed (Figs. 1 and 2). The chromatographic locations of deconvoluted components are shown in zoom in Fig. 1. Taking the sample whose concentration is 500 μM as an example, the deconvoluted spectra of metabolite derivatives are compared with the standard spectra in the libraries (Fig. 2).

When the concentration was 500 μM, AnalyzerPro deconvoluted 8 components (Fig. 1a). Fig. 2 showed that the deconvoluted mass spectra of 4 components and 1 *n*-alkane coincide with the standard spectra and all their match factors are greater than 850 for uracil, 864 for nicotinic acid, 915 for glutaric acid, 918 for *n*-pentadecane and 903 for citramalic acid. The remaining 3 components are extraneous, aberrant peaks (false positives).

When the solution concentration decreased to 350 μM (Fig. 1b), 6 components were detected, and 3 metabolite derivatives (uracil, glutaric acid and citramalic acid) and *n*-pentadecane were identified. The previously shown, validated peak for nicotinic acid in Solution 1 was not detected (false negative). The match factors of metabolites detected decreased, and that of glutaric acid decreased to 833 (i.e. less than 850). The remaining 2 components are extraneous, aberrant peaks (false positives).

Then, when the solution concentration decreased to 150 μM (Fig. 1c), 5 components were detected. The result for Solution 3 was less accurate than for Solution 2 in that the ions contained in deconvoluted spectra were inaccurate, which meant that the qualitative result was not adequate, although the components were deconvoluted at the correct retention time. The match factors of all 3 identified metabolites were less than 850 (i.e. 673 for uracil, 293 for glutaric acid and 819 for citramalic acid). Only the deconvoluted mass spectrum for *n*-pentadecane was correct (match factor 908). The remaining component is an extraneous, aberrant peak (false positive).

Finally, when the concentration of the solution decreased to 50 μM (Fig. 1d), the accuracy was the least, as only *n*-pentadecane was detected and identified correctly, and no metabolites were detected.

From these results, we observed that the deconvoluted results strongly depend on the concentration of metabolites in the sample. When the concentration is comparatively high, the software is liable to produce false positives, but when the concentration is comparatively low, it produces more false negatives and the quality of deconvoluted mass spectra decreases, even though they are detected. The same trend can be observed in the complete data set (Table 2). With the concentration decreasing from 500 μM to 50 μM, the number of components with match factor over 850, acquired by ChromaTOF, decreased from 37 to 14; that by AMDIS from 32 to 8; that by AnalyzerPro from 28 to 5 – which indicated that the deconvoluted result strongly depends on the concentration.

However, many metabolites are usually present at a relatively low concentration in biological samples, and, from a biological point of view, metabolites present in high concentrations are not necessarily more important than those present at low concentrations, so the problem of how to identify the components with low concentrations in biological samples remains a challenge for us.

3.2. Analysis of repeatability

An essential factor in assessing the quality of software is the repeatability of results. Ideally, when the software processes the data from different technical replicates of

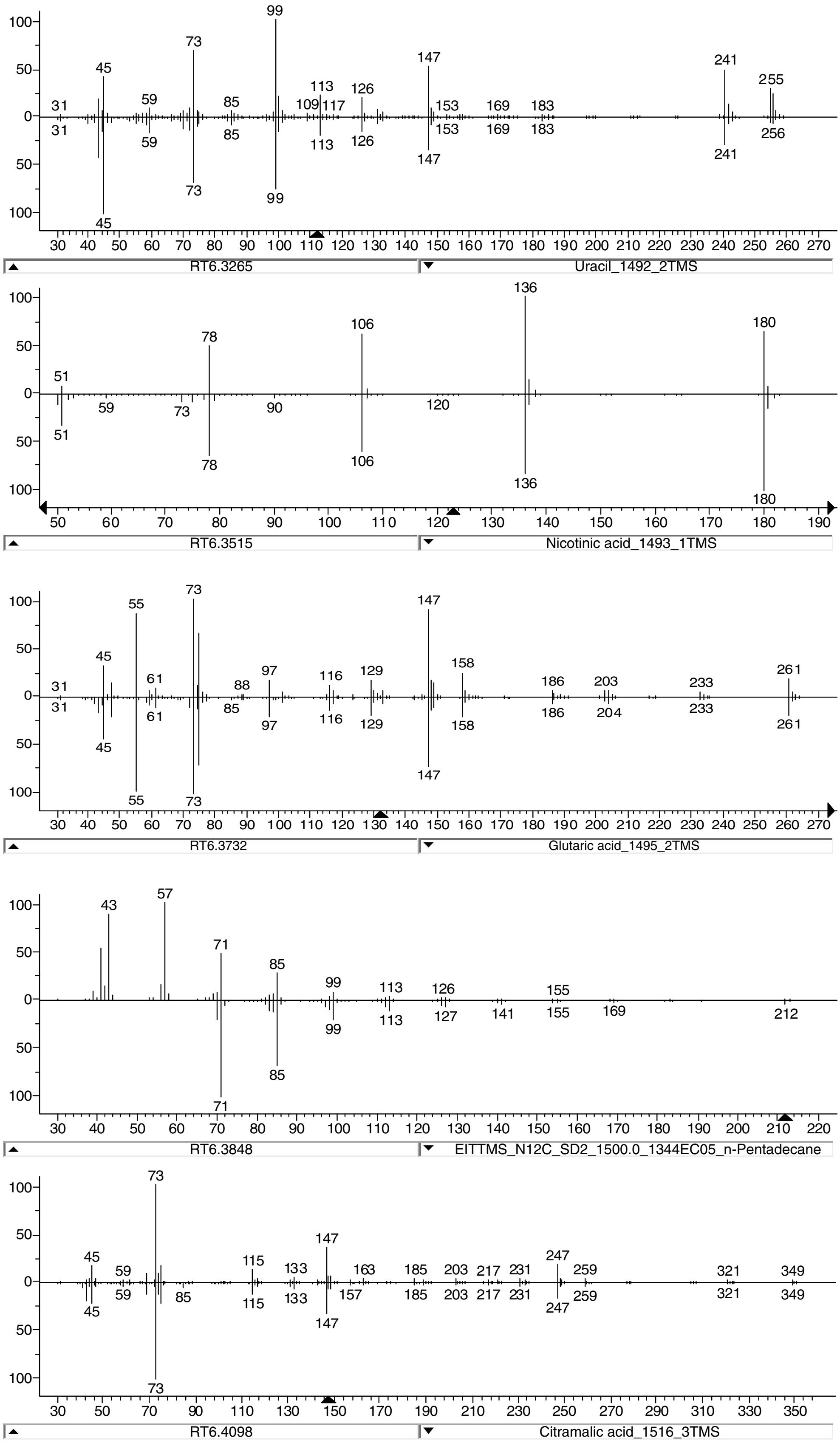

Figure 2. Deconvoluted spectra of metabolites whose concentration is 500 μM, compared with standard spectra in library.

the same sample, the results provided by software should remain constant, because the sample contains the same metabolites. To test the repeatability of the software, accurate sample-preparation and chromatography processes are essential to produce repeatable results. Previous studies [12,28] have shown that the reliability of this experimental method is very high.

We considered another technological challenge encountered in metabolomics – dynamic range – as there are large differences in the concentrations of the different metabolites present in a metabolomics data set. In this test, Solution 5, with the wide variation of metabolite concentrations, was prepared and analyzed three times to test the repeatability of software packages. The TICs of the 3 replicates are shown in Fig. 3.

The retention-time window of 416–423 s in TIC of Solution 5 (see Fig. 3) included 6 metabolite derivatives (fructose, cis-aconitic acid, fructose, glucose, citric acid and glucose), and the results from each software package are shown in Table 3.

Although we expected that the 6 metabolite derivatives in the 3 replicates to be deconvoluted correctly and identically, the numbers of deconvoluted components and spectra for the replicate analyses differ from each other

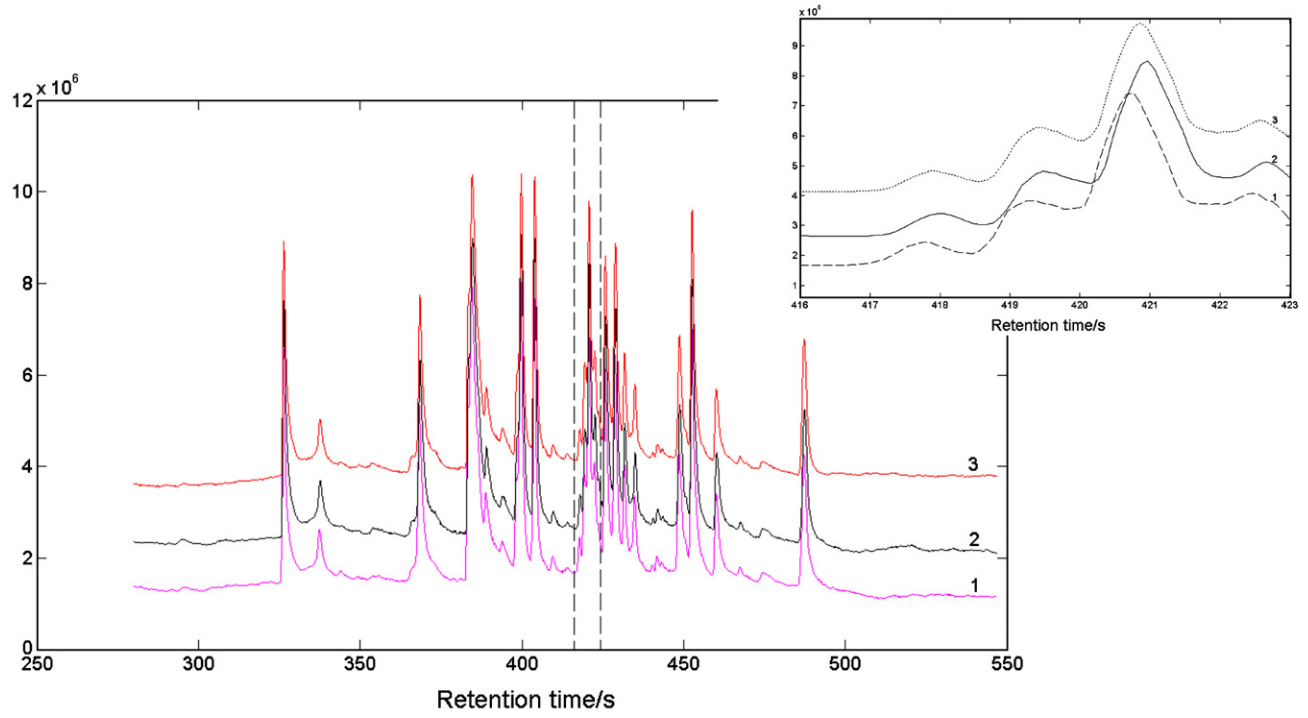

Figure 3. TICs of 3 replicates for Solution 5 and the retention-time window of 416–423 s.

Table 3. Results from fragment of 416–423 s in replicates using 3 software packages

Software	Replicate	Number of components deconvoluted	Metabolite derivatives					
			Fructose	*cis-Aconitic acid*	*Fructose*	*Glucose*	*Citric acid*	*Glucose*
ChromaTOF	1	10		○	○		○	
	2	10	○	○	○			
	3	9	○	○	○			
AMDIS	1	43		○	○		○	
	2	31		○	○			○
	3	44		○	○		○	
AnalyzerPro	1	6			○		○	○
	2	6			×		○	
	3	6			○		○	○

○ = Detected metabolite derivative without wrong spectrum; × = Undetected metabolite derivative; Empty = Detected metabolite derivative with right spectrum.

Table 4. Comparison of results generated by 3 software packages for Solution 5 in replicated runs

Data		Replicate 1	Replicate 2	Replicate 3
Number of components deconvoluted	ChromaTOF	149	162	159
	AMDIS	782	720	995
	AnalyzerPro	44	42	45
Number of metabolite derivatives undetected	ChromaTOF	0	0	0
	AMDIS	0	0	0
	AnalyzerPro	16	19	18
Number of metabolite-derivative spectra deconvoluted correctly	ChromaTOF	30	27	30
	AMDIS	26	26	23
	AnalyzerPro	19	18	18

(Table 3). AMDIS and ChromaTOF detected all 6 metabolite derivatives with many false positives, and inconsistent and unsatisfactory spectra. AnalyzerPro produced fewer false positives, but, in the second replicate, fructose was not deconvoluted (false negative). A similar tendency can be observed for all the data (Table 4).

From the results in Tables 3 and 4, the repeatability of the software packages was not as high as expected. Of course, this problem may be attributed to two factors, the software and the experiment.

3.3. *Effect of software parameters*

To understand better those parameters that affect the deconvolution accuracy, different parameter values were tested in the 3 software packages. When testing one parameter, the other parameters remained constant as shown in Section 2.3. The results from the 3 software packages using different parameter settings are shown in Tables 5–7.

As shown in Table 5, the parameters employed in AMDIS had no impact on the number of undetected metabolite derivatives (i.e. it deconvoluted all metabolite derivatives in the sample, though with a large number of false positives). Meanwhile, as parameters adjacent peak subtraction, resolution and sensitivity increased from low to high, the number of deconvoluted components increased. As the shape requirements were changed from low to high, the number of deconvoluted components

Table 5. Results from AMDIS with different settings

Parameter	Parameter value	Number of components deconvoluted	Number of undetected metabolites out of 51 expected metabolite derivatives	Number of spectra deconvoluted correctly out of 51 expected metabolite derivatives
Component width	3	480	0	17
	6	637	0	23
	8	681	0	24
	9	701	0	27
	10	705	0	26
	12	720	0	26
	15	737	0	25
	20	777	0	24
Adjacent peak substraction	0	690	0	26
	1	720	0	26
	2	720	0	24
Resolution	low	571	0	26
	medium	720	0	26
	high	996	0	29
Sensitivity	very low	210	0	26
	low	392	0	27
	medium	720	0	26
	high	1361	0	27
	very high	2131	0	26
Shape requirements	low	728	0	27
	medium	720	0	26
	high	647	0	19

Table 6. Results from AnalyzerPro with different settings

Parameter	Parameter value	Number of components deconvoluted	Number of undetected metabolites out of 51 expected metabolite derivatives	Number of spectra deconvoluted correctly out of 51 expected metabolite derivatives
Component width (min)	0.001	44	18	20
	0.01	44	18	20
	0.02	42	19	18
	0.03	34	23	13
	0.04	27	26	12
	0.05	21	31	11
	0.10	4	47	0
Minimum masses	3	63	11	18
	6	42	19	18
	9	35	21	18
	12	32	23	18
Resolution	minimum	51	13	19
	low	42	19	18
	high	42	19	18
	maximum	42	19	18
Scan windows	1	46	16	17
	2	42	19	18
	5	40	19	18
	9	37	20	17
Smoothing	1	42	19	18
	5	69	9	20
	9	82	7	20
	15	82	8	20
Area threshold	50	43	18	18
	500	42	19	18
	1000	36	20	18
	2000	33	22	15
	3000	30	24	15
High threshold	20	42	18	19
	200	42	19	18
	1000	38	20	18
	3000	32	23	15
	4000	30	24	14

decreased. So we recommend parameter settings in AMDIS that will deconvolute a lower number of components; of course, that is sometimes likely to sacrifice to some extent the accuracy of deconvoluted spectra of metabolite derivatives. For example, when the resolution value was set at low or medium, the number of deconvoluted components is apparently smaller than when the value was set at high (see Table 5), but, unfortunately, the number of correctly deconvoluted spectra also decreased.

From the results shown in Table 6, when changing parameters, including minimum mass, resolution, scan windows, area threshold and height threshold, the number of deconvoluted components decreased and the number of undetected metabolite derivatives increased, but there was no apparent effect on the number of correctly deconvoluted spectra. With the value of parameter smoothing increasing, the number of components increased and the number of undetected metabolite derivatives decreased (i.e. false positives increased and false negatives decreased). At first, the number of correctly deconvoluted spectra increased, with a maximum of 5, so the parameter values suggested by software provider (see Section 2.3) were reasonable.

The results shown in Table 7 for ChromaTOF indicated that altering the parameters for baseline, smoothing and the S/N ratio have little impact on the deconvolution result.

To summarize, the results of the 3 software packages indicated that component width was the chief determinant of the deconvolution result. The influence of other parameters (e.g., smoothing, baseline, scan windows and resolution) was relatively weak. The peak widths in data for Solution 5 were 0.02–0.04 min or 9–15 scan points. We found that the closer the setting of the component

Table 7. Results from ChromaTOF with different settings

Parameter	Parameter value	Number of components deconvoluted	Number of undetected metabolites out of 51 expected metabolite derivatives	Number of spectra deconvoluted correctly out of 51 expected metabolite derivatives
Component width (s)	0.4	456	0	30
	1.2	220	0	31
	2.5	162	0	27
	4.0	131	1	26
	8.0	86	4	23
	2	180	0	26
S/N	5	179	0	26
	10	162	0	27
	15	151	1	27
	0	161	0	25
Baseline	0.5	165	0	25
	1	162	0	27
	1	149	1	28
	3	162	0	27
Smoothing	5	156	0	29
	7	161	0	25
	9	150	1	24
	11	143	1	25
	15	141	2	26

width is to the true value, the greater is the number of spectra of metabolite derivatives that are deconvoluted correctly (Tables 5–7).

The default in the 3 software packages can be used when we are not sure which value should be set. To obtain an optimal result, it is most important to estimate the component width as accurately as possible. The results slightly depended on the type of data file (e.g., noise or peak overlap), which could be found in data acquired by different ramps (data not shown). The peak widths observed in metabolomic studies are generally variable and therefore using one peak width can be detrimental.

We recommend that the parameter shape should be set at low in AMDIS, because the chromatographic peak shape for derivatized metabolite peaks are not consistent because of the sample complexity.

4. Discussion

We have evaluated three commercially or freely available software packages (ChromaTOF, AnalyzerPro and AMDIS) for the analysis of data from metabolite mixtures analyzed with GC-TOF-MS. The aim of this research was to evaluate and to compare the applicability of existing software packages, to highlight the requirements and the difficulties, and to promote discussion on possible solutions for metabolomics chromatographic data.

Based on all results above, the 3 different software packages each have advantages and limitations.

One particular feature of the present ChromaTOF and AMDIS deconvolution-software packages is that they tend to generate artifactual components corresponding to noise (as judged by the mass spectrum and TIC chromatogram) and to produce duplicate or multiple peak assignments that (again from the mass spectra) clearly correspond to a single chromatographic peak and chemical entity. Such artifacts can account for 10–20% of the peaks in the chromatogram for ChromaTOF and 70–80% for AMDIS, although they adequately detected closely co-eluting components. AnalyzerPro results showed the detection of fewer false positives, though also with a greater number of false negatives. Some metabolites present in the sample could not be detected. It should be noted that up to the release of this paper, there has been a new revision to the AnalyzerPro algorithm to address the number of false negatives without increasing the false positives.

The results of the 3 software packages strongly depended on the concentration of sample. An attempt should be made to reconstruct ''pure-component'' spectra from complex TIC chromatograms, even when components are present at trace levels. For this purpose, observed chromatographic behavior, along with a range of noise-reduction methods, are expected to work.

Repeatability of all software is unsatisfactory. To improve repeatability, we should try to modify the algorithm or to find a new algorithm for deconvolution. A standardized protocol for sample preparation is necessary.

Although all results depend on operator-set software parameters, there is no one set of values in the 3 software packages that will be successful for the deconvolution of all metabolites. When the accurate component width was provided, the optimal analysis result could be obtained. The other parameters only slightly influenced conclusions based on the data and, subsequently, the reported numbers of identifications.

Since there were differences between the programs in both the number of confident identifications and the components detected, there was no specific indication that any of the programs was superior.

Combination of the results acquired by the 3 software packages could circumvent problems in a complementary way, and may improve the reliability of results.

As there is no better software package available presently, from the results of our tests, we suggest that the user should choose the deconvolution software for metabolomics study according to their experimental objectives. If you prefer obtaining an accurate number of the metabolites in samples from mass spectra, we recommend ChromaTOF and AnalyzerPro. If you prefer accurate mass spectra, AMDIS and ChromaTOF are better choices.

As far as speed is concerned, although automatic software packages can also be applied with some degree of success, they are still a fairly slow for the flood of data from metabolomics. AnalyzerPro can handle multiple data files simultaneously to speed up the process. In addition, AnalyzerPro and ChromaTOF possess a comparatively friendly user interface, and the display, the input and the output of data and figures are more convenient than those of AMDIS.

Another important point is that the use of ChromaTOF is restricted to data with vendor-instrument-specific file formats, while AMDIS and AnalyzerPro can process multiple vendors' instrument data through a single user interface, so consistent data analysis and presentation from different instruments help with the development of standard operating procedures (SOPs) and client reports. This is very useful for metabolomics development, because metabolomics is a strategy increasingly being applied and requires many researchers to participate, so the data from different instruments and laboratories may be expected to be processed with the same software packages.

Generally, none of these 3 software packages has provided a comprehensive solution to meet the challenges or the needs for the development of metabolomics. More efficient, automated, flexible and reliable data-handling systems are required. Future developments in this area are vital for metabolomics to progress. It is necessary to find new algorithms and to write better software that can avoid false positives and false negatives, and that can deconvolute low-concentration components from high noise and background. Availability of a vendor-independent data-processing software pipeline that is modular and flexible enough to incorporate new algorithms and that is expandable to other types of MS data (e.g., GC–MS and CE–MS) could significantly boost progress in metabolomics.

Acknowledgements

Spectralworks Ltd is thanked for the trial use of AnalyzerPro software, in particular, John Moncur for useful suggestions. The AMDIS program is downloaded free; details may be obtained from the website http://chemdata.nist.gov/mass-spc/amdis/. ChromaTOF software is available as part of the LECO Pegasus III TOF-MS instrument. This work was supported by a China Partnering Award grant from the UK Biotechnology and Biological Sciences Research Council (BBSRC) (Grant PA 1479). H.L. also thanks National Natural Science Foundation of China and Hunan Provincial Natural Science Foundation for support of the projects (No. 20745005 and No. 07JJ4004). D.B.K. is EPSRC/RSC Research Chair in Bioanalytical Science.

References

[1] W.B. Dunn, N.J.C. Bailey, H.E. Johnson, Analyst (Cambridge, U.K.) 130 (2005) 606.

[2] O. Fiehn, Plant Mol. Biol. 48 (2002) 155.

[3] S.G. Villas-Bôas, U. Roessner, M.A.E. Hansen, J. Smedsgaard, J. Nielsen (Editors), Metabolome Analysis: An Introduction, John Wiley and Sons Inc., New York, USA, 2007.

[4] W.B. Dunn, D.I. Ellis, Trends Anal. Chem. 24 (2005) 285.

[5] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Trends Biotechnol. 22 (2004) 245.

[6] D.B. Kell, Curr. Opin. Microbiol. 7 (2004) 296.

[7] D.B. Kell, M. Brown, H.M. Davey, W.B. Dunn, I. Spasic, S.G. Oliver, Nat. Rev. Microbiol. 3 (2005) 557.

[8] J. Forster, I. Famili, P. Fu, B.O. Palsson, J. Nielsen, Genome Res. 13 (2003) 244.

[9] N.C. Duarte, S.A. Becker, N. Jamshidi, I. Thiele, M.L. Mo, T.D. Vo, R. Srivas, B.O. Palsson, Proc. Natl. Acad. Sci. USA 104 (2007) 1777.

[10] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, BMC Genomics 7 (2006) 142.

[11] K. Hollywood, D.R. Brison, R. Goodacre, Proteomics 6 (2006) 4716.

[12] S. O'Hagan, W.B. Dunn, M. Brown, J.D. Knowles, D.B. Kell, Anal. Chem. 77 (2005) 290.

[13] S. O'Hagan, W.B. Dunn, J.D. Knowles, D. Broadhurst, R. Williams, J.J. Ashworth, M. Cameron, D.B. Kell, Anal. Chem. 79 (2007) 464.

[14] G.G. Harrigan, R. Goodacre (Editors), Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis, Kluwer Academic Publishers., London, UK, 2003.

[15] P.J. Gemperline, J. Chem. Info. Comp. Sci. 24 (1984) 207.

[16] E.J. Karjalainen, Chemometrics Intelligent Lab. Syst. 7 (1989) 31.

[17] O.M. Kvalheim, Y.Z. Liang, Anal. Chem. 64 (1992) 936.

[18] Y.Z. Liang, O.M. Kvalheim, H.R. Keller, D.L. Massart, P. Kiechle, F. Erni, Anal. Chem. 64 (1992) 946.

[19] E.R. Malinowski, J. Chemometrics 10 (1996) 273.

[20] R. Manne, B.-V. Grande, Chemometrics Intelligent Lab. Syst. 50 (2000) 35.

[21] Y.Z. Liang, O.M. Kvalheim, Fresenius' J. Anal. Chem. 370 (2001) 694.

[22] J. Lisec, N. Schauer, J. Kopka, L. Willmitzer, A.R. Fernie, Nat. Protocols 1 (2006) 387.

[23] P. Mendes, Brief. Bioinformatics 7 (2006) 127.

[24] U. Roessner, C. Wagner, J. Kopka, R.N. Trethewey, L. Willmitzer, Plant J. 23 (2000) 131.

[25] J.M. Halket, D. Waterman, A.M. Przyborowska, R.K.P. Patel, P.D. Fraser, P.M. Bramley, J. Exp. Bot. 56 (2005) 219.

[26] B.Y. Li, Y. Hu, Y.Z. Liang, L.F. Huang, C.J. Xu, P.S. Xie, J. Sep. Sci. 27 (2004) 581.

[27] B.Y. Li, Y.Z. Liang, P.S. Xie, R.Q. Yu, Chin. J. Anal. Chem. 31 (2003) 799.

[28] W.B. Dunn, D.I. Broadhurst, S.M. Deepak, M.H. Buch, G. McDowell, G. Spasic, D.I. Ellis, N. Brooks, L. Neyses, D.B. Kell, Metabolomics 3 (2007) 413.