DOI: 10.1007/s11306-005-4432-7

A non-directed approach to the differential analysis of multiple LC–MS-derived metabolic profiles

O. Vorst[a,*], C. H. R. de Vos[a], A. Lommen[a,d], R. V. Staps[a], R. G. F. Visser[b], R. J. Bino[a,c], and R. D. Hall[a,e]

[a]*Plant Research International, PO Box 16, 6700 AA, Wageningen, The Netherlands*
[b]*Laboratory of Plant Breeding, PO Box 386, 6700 AJ, Wageningen, The Netherlands*
[c]*Laboratory of Plant Physiology, Wageningen University, Arboretumlaan 4, 6703 BD, Wageningen, The Netherlands*
[d]*RIKILT-Institute of Food Safety, PO Box 230, 6700 AE, Wageningen, The Netherlands*
[e]*Centre for BioSystems Genomics (CBSG), PO Box 98, 6700 AB, Wageningen, The Netherlands*

Received 10 February 2005; accepted 17 March 2005

An essential element of any strategy for non-targeted metabolomics analysis of complex biological extracts is the capacity to perform comparisons between large numbers of samples. As the most widely used technologies are all based on mass spectrometry (e.g. GCMS, LCMS), this entails that we must be able to compare reliably and (semi)automatically large series of chromatographic mass spectra from which compositional differences are to be extracted in a statistically justifiable manner. In this paper we describe a novel approach for the extraction of relevant information from multiple full-scan metabolic profiles derived from LC–MS analyses. Specifically-designed software has made it possible to combine all mass peaks on the basis of retention time and m/z values only, without prior identification, to produce a data matrix output which can then be used for multivariate statistical analysis. To demonstrate the capacity of this approach, aqueous methanol extracts from potato tuber tissues of eight contrasting genotypes, harvested at two developmental stages have been used. Our results showed that it is possible to discover reproducibly discriminatory mass peaks related both to the genetic origin of the material as well as the developmental stage at which it was harvested. In addition the limitations of the approach are explored by a careful evaluation of the alignment quality.

KEY WORDS: metabolomics; spectral alignment; data mining; potato; liquid chromatography–mass spectrometry; multivariate analysis.

1. Introduction

Research efforts in the field of plant functional genomics are focussed on integrating molecular data describing the process of gene expression at its different levels (mRNA, protein, metabolite) in order to understand better the resulting phenotypic characteristics (e.g. Goossens *et al.*, 2003; Hirai *et al.*, 2004; Weckwerth *et al.*, 2004). The parallel analysis of mRNAs (Aharoni and Vorst, 2002; Meyers *et al.*, 2004; Schnable *et al.*, 2004), and of proteins (Gallardo *et al.*, 2002; Watson *et al.*, 2003; Cánovas *et al.*, 2004) has become an established procedure. However, many functional genomics approaches miss the analysis of the functional bioactive components: the metabolites. The non-targeted analysis of the metabolite complement of a cell, generally designated as metabolomics, is still somewhat less developed (Fiehn, 2002; Hall *et al.*, 2002; Sumner *et al.*, 2003; Weckwerth, 2003; Bino *et al.*, 2004). Currently, gas chromatography–mass spectrometry (GC–MS; Fiehn *et al.*, 2000; Roessner *et al.*, 2000, 2001a; Wagner *et al.*, 2003) and to a lesser extent liquid chromatography–mass spectrometry (LC–MS; Von Roepenack-Lahaye *et al.*, 2004) are the analytical techniques most often applied to metabolomics analysis today. Nevertheless, the use of these technologies is often limited due to difficulties that arise from the sheer size and complexity of the datasets obtained. Combining the results obtained from several biological samples into a single differential analysis is an arduous task that requires the matching of peaks representing the same compound over several chromatograms. When combining mRNA expression profiles originating from microarray experiments, an unequivocal identity can be derived from the array position. In contrast, a direct linkage of complete metabolite profiles based on peak identities is not feasible as many chromatographic peaks can only be identified with a limited degree of certainty. Here we explore an alternative approach in which peaks are first linked based on both retention time and m/z characteristics, prior to any identification of differential mass signals.

Due to its relatively robust chromatography and compound separation efficiency, resulting in reproducible retention times of hundreds of mass peaks, together with the availability of reference compound libraries, GC-(TOF)-MS of derivatized extracts is at present generally preferred over LC–MS in metabolomic studies

* To whom correspondence should be addressed.
E-mail: oscar.vorst@wur.nl

(Fiehn *et al.*, 2000; Roessner *et al.*, 2001a, b; Wagner *et al.*, 2003; Fernie *et al.*, 2004). However, GC–MS is less applicable to semi-polar compounds among which are major classes of plant (secondary) metabolites including flavonoids, (glyco-)alkaloids, glucosinolates and saponins. Recent advances in techniques for improving resolution in LC by using capillary electrophoresis (Soga *et al.*, 2002), hydrophilic interaction columns (Tolstikov and Fiehn, 2002) and monolithic columns (Tolstikov *et al.*, 2003) demonstrate a high potential for LC–MS complementing GC–MS in unravelling metabolic profiles. Recently, capillary LC combined with high resolution QTOF-MS has been successfully applied to the analysis of *Arabidopsis* secondary metabolites (Von Roepenack-Lahaye *et al.*, 2004). However, as a result of non-linear shifts in retention times and other technical limitations inherent to liquid chromatography it is not feasible to do a reliable comparative analysis of LC–MS profiles without prior manipulations (deconvolution, re-alignment etc).

The present study was carried out using potato tubers of eight selected genotypes of the diploid C × E population which is a genetically well characterized, highly diverse breeding population (Celis Gamboa, 2002). Extracts of tubers harvested at two different moments in the growing season were subjected to high-resolution reversed phase LC-QTOF MS analysis. In order to align all the mass chromatograms obtained and to detect differentially accumulating metabolites in the tubers in an unsupervised way, we used the metAlign software, and subsequently performed principal component analyses (PCA) on the resulting data matrix (aligned peaks × samples). Based on the aligned mass profiles the tubers were separated by PCA into two groups, corresponding to the two different moments of harvest. Furthermore, our analysis pointed to those peaks significantly correlating with these phenotypic differences. This is the first report employing the metAlign software on large LC–MS-derived metabolomics datasets.

2. Materials and methods

2.1. Plant material

Field-grown potato (*Solanum tuberosum* L.) tubers were harvested from eight different genotypes (155, 276, 668, 673, 674, 697, 732 and 738) of the C × E breeding population (Celis Gamboa, 2002) at two time points: an early time point, adjusted to the earliness of each genotype, aimed at obtaining tubers at the same developmental stage (mid-way), and the final harvest time point identical for all genotypes. The latter was at the end of the growing season, after tuber maturation and herbicide foliage spraying. Each sample, comprising all tubers from two plants of the same genotype, was frozen in liquid nitrogen and stored at −70 °C.

2.2. Sample preparation

About 150 g of frozen tubers (including material from all parts of the tubers) were ground to a fine powder in liquid nitrogen. Extracts of 0.5 g frozen powder in 2 mL 62.5% methanol, 0.125% formic acid in water at 0°C were prepared in duplicate. After immediate mixing, the extracts were sonicated for 10 min, spun down (10 min at 1000 × *g*) and filtered through a 0.2-μm inorganic Anotop 10 membrane filter (Whatman 6809–1022).

2.3. HPLC–MS analysis

Samples were automatically injected and separated using an Alliance 2795 HT system (Waters Corporation) equipped with a Luna C_{18}-reversed phase column (150 × 2.1 mm, 3 μm; Phenomenex, CA) at 40°C and a gradient from 5 to 50% (see figure 1c) acetonitrile acidified with 0.1% formic acid at a flow rate of 0.2 mL/min. Samples were run in random order in a single batch. Compounds eluting from the column were detected on-line over a period of 47 min, first by a Waters 996 photodiode array detector at 210–600 nm and then by a Q-TOF Ultima MS (Waters). Electron Spray Ionisation (ESI), in positive mode, and MS settings were initially optimized using direct infusion of a potato extract diluted in 50% acetonitrile with 0.1% formic acid, resulting in the following conditions being used during LC–MS runs: desolvation temperature of 300°C with a nitrogen gas flow of 500 L/h, capillary spray at 3 kV, source temperature of 120°C, cone voltage at 35 V with 50 L/h nitrogen gas flow, collision energy at 5 eV. Ions in the *m*/*z* range 100–1500 were detected using a scan time of 900 ms and an inter-scan delay of 100 ms. The MS was calibrated using 0.05% phosphoric acid in 50% acetonitrile. Leucine enkaphalin, detected on-line through a separate ESI-interface every 10 s, was used as a lock mass for exact mass measurements (Wolff *et al.*, 2001). MassLynx software version 4.0 (Waters) was used to control all instruments and for calculation of the accurate masses.

2.4. Data handling and alignment

Chromatographic data generated in MassLynx format were directly imported into the metAlign software (http://www.metAlign.nl). After optimizing the settings according to the specific chromatographic conditions, this software is able to compare samples based on the ions detected in an unbiased and unsupervised manner by performing the following steps: (a) data smoothing by digital filters related to the average peak width, (b) estimation of local noise as a function of retention time and ion trace, (c) baseline correction of ion traces and introduction of a threshold to obtain noise reduc-

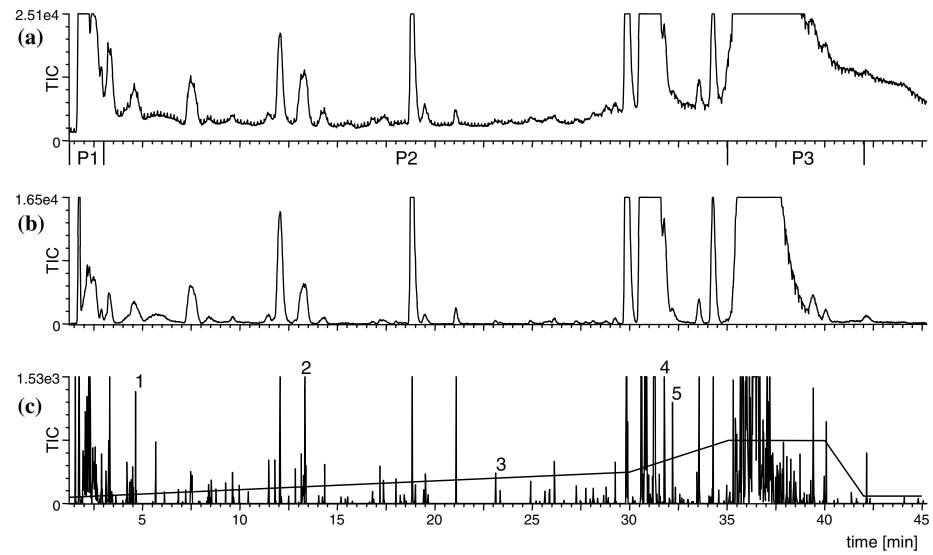


Figure 1. Typical LC–MS chromatogram of a tuber extract (genotype 674, early harvest). Shown is (a) the total ion count (TIC) in positive mode, (b) the same after metAlign-assisted baseline correction and (c) peak deconvolution. In (c) the gradient used is indicated as % acetonitrile; in (a) the three chromatographic phases discerned are indicated: flow-through phase (P1; scan 67–156), gradient phase (P2; 157–1830) and washing phase (P3; 1831–2214). Identified peaks in (c) are labelled as follows: 1, phenylalanine, 2, chlorogenic acid, 3, rutin, 4, α-chaconine, 5, α-solanine.

tion, (d) calculation and storage of peak maximum amplitudes, (e) between-chromatogram alignment using high S/N peaks common to all chromatograms (‘landmark peaks’), (f) iterative fine alignment by including an increasing number of landmark peaks with lower S/N. For each mass trace the noise is estimated as a function of time. The background elimination algorithm applied is not a simple threshold initiated data reduction but utilizes a baseline shape independent series of linear corrections on individual mass traces (Lommen *et al.*, 1998; Noteborn *et al.*, 2000). All mass peaks in all datasets are used in an alignment algorithm to obtain an ordered data matrix (‘aligned peaks’ vs. samples) which can be exported in a format compatible with most multivariate software packages. As an alternative, a univariate approach to a simple two group differential problem (*t*-test) can be undertaken by metAlign resulting in the exporting of differential datasets to MS-data formats for direct visual validation.

For multivariate statistical analysis the potato data matrix (aligned peaks × samples) of the normalized peak intensities resulting after alignment step (f) was used. Aligned peaks represented in <4 samples were filtered out. Peaks absent in certain datasets within this matrix were automatically awarded a value 3 times the estimated local noise. The generated data, in csv-file format, were then log-transformed and imported into GeneMaths software (Applied Maths, Sint-Martens-Latem, Belgium) for further analysis. PCA was performed on normalized data (mean centred with respect to the aligned peaks and samples). The same data matrix was used to calculate the standard deviation (SD) of intensities per aligned mass peak (row) as follows. Only mass peaks present in both duplicate analyses were included. To compensate for compositional differences between the samples, each duplicate peak pair was normalized using their average intensity. Finally, the SD was calculated for each aligned mass peak (DF = $N/2$).

2.5. Accurate mass calculation

Calculation of accurate masses was done using the ‘Accurate Mass Measure’ tool of MassLynx version 4.0. Settings were: Background subtraction: Polynomial order = 5, Below curve (%) = 33; Smooth window = 1, Number of smooths = 2, mean; Min peak width at half height = 4, Centroid top = 80; TOF constants: Resolution = 10,000, Np Multiplier = 0.7; Lock mass set at 556.2771, 3 scans averaged. Resulting files were converted from MassLynx- into an ASCII-format using the DataBridge utility of MassLynx.

The accurate masses obtained were integrated with metAlign results by using a Perl script that takes the retention time matrix of metAlign (End_result_retentions.csv) and the accurate mass data of each run (ASCII-format) as input. It produces two matrices, in which the retention times are replaced by the calculated accurate masses and peak intensities, respectively. To calculate each accurate mass, the most abundant mass peak was selected from within the 1 Da mass bin (the accuracy resulting from metAlign) in the scan number (derived from the retention time) according to the metAlign output. To enhance the precision, average accurate masses were calculated from three consecutive scans. The total intensity (recorded as ion counts) of these three peaks was stored in a separate matrix for evaluation purposes.

3. Results and discussion

3.1. LC–MS based metabolic profiles

Acidified aqueous methanol extracts of tubers from eight different potato genotypes harvested at two time points were analyzed using reversed phase HPLC-QTOF MS. A typical complete chromatogram, showing the total ion count (TIC), is given in figure 1a. Three phases in the chromatogram are discerned: a 'flow-through' phase (P1; scan 67–156), a 'gradient' phase (P2; scan 157–1830) and a 'washing' phase (P3; scan 1831–2214). The gradient phase represents the chromatogram proper, to which data interpretation usually would be limited. However, to also evaluate the validity of our approach under less favourable conditions, the chromatographically sub-optimal 'flow-through' and 'washing' phase were included in the present analyses.

After metAlign-assisted baseline correction (figure 1b) and subsequent peak deconvolution (figure 1c), the number of mass peaks having an intensity of at least 10 times the estimated local noise ranged from 1400 to 2000 per analysis. The frequency distribution of peak intensities is shown in figure 2. A large fraction of the extracted peaks has low (<16 times local noise) though significant peak intensities. To evaluate the reproducibility of the extraction and chromatographic procedures applied, each analysis was repeated once from the same starting material (frozen powder) for all samples studied. This allowed for the calculation of the SD per normalized mass signal over the samples, using the matrix output of metAlign. A total of 1175 mass peaks were present in all of the 28 duplicate analyses and the average SD in the intensity of these peaks was 16.1 ± 7.4%. For the 1840 peaks represented in at least 22 of the analyses the average SD was 18.9 ± 10.1%.

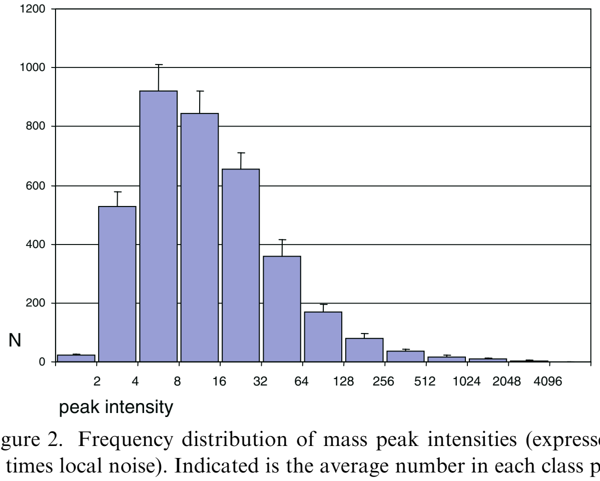


Figure 2. Frequency distribution of mass peak intensities (expressed as times local noise). Indicated is the average number in each class per LC–MS run. Error bars indicate the SD over 30 LC–MS runs analyzed. The average number of peaks per run was 3645 ± 326.

3.2. Quality of the alignment

The quality of the resulting multiple alignment of a diverse chromatographic dataset, as produced by metAlign, is difficult to assess in the absence of any generally accepted method as a bench mark. The manual alignment of hundreds of peaks that could serve as a substitute method is a virtually impossible task. Hence an alternative approach has been used to evaluate the quality of the results obtained, based on the fact that metAlign rounds off accurate masses in its alignment as peaks are binned within a 1 Da window (ranging from −0.35 to +0.65 Da). A close analysis of the accurate masses of the aligned peaks can assist in gaining insight into their homogeneity and hence into the quality of the alignment. The accurate mass of the major peak present in each 1 Da mass window was extracted from processed chromatograms with lock-mass-corrected accurate mass measurements (AFAMM), as described in the Materials and methods section. Data from three subsequent scans were used in order to obtain a better estimation of the exact masses of the peaks.

A drawback of an approach that restricts itself to retention time and molar mass as the basis for combining mass peaks, is that it relies upon the assumption that the samples under study are similar and hence possess a reasonably comparable metabolic profile. The absence of a considerable number of compounds in one or more of the samples might destabilize the alignment, and force the method employed to combine mass peaks that represent distinct compounds (hyper-alignment). On the other hand, an algorithm that is too reluctant in aligning peaks will unintendedly fail to recognize cognate peaks (hypo-alignment). These two major error types can be envisaged potentially to disturb the proper alignment of the chromatographic peaks. Hyper-alignment will result in peaks in different chromatograms that actually represent distinct compounds being erroneously combined as they have a similar retention time and are within the same mass bin. Hypo-alignment will ensure that peaks representing the same compound are not recognized as such and do not end up in the same aligned peak (row in the matrix). Both hyper- and hypo-alignment will lead to faulty conclusions in a difference analysis, as these types of mistakes are not properly dealt with in standard statistical analyses.

3.2.1. Assessing hyper-alignment

Hyper-alignment in the metAlign-derived data matrix (in our example limited to peaks of at least 10× local noise) was studied using the variation of accurate masses within the aligned peaks (rows). To avoid the inclusion of imprecise accurate mass estimates, masses derived from less than 100 ion counts (three consecutive scans) were excluded from the analysis. The resulting frequency distribution of the SDs of the aligned peaks (figure 3) is

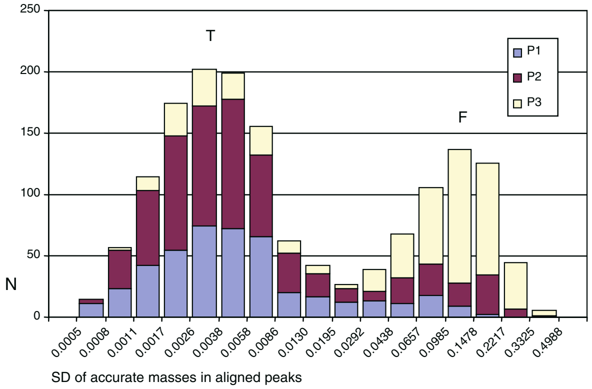


Figure 3. Frequency of SD of accurate masses in single aligned peaks. Analysis was limited to signals of at least 10× local noise (metAlign) and having their accurate mass calculation based on at least 100 ion counts (over 3 scans). The histogram is based on the 1575 aligned peaks represented in at least 3 samples. Aligned peaks were classified as flow-through (P1), gradient (P2) and washing phase peaks (P3) based on their elution time (see figure 1a).

clearly bimodal in nature, with a divide at 0.029 Da. The class of aligned peaks (T) with a low SD (< 0.029 Da) corresponds to combined peaks that appear homogenous in nature as they share the same major compound within the cognate mass bin for all peaks combined. The second class (F) (SD ≥ 0.029 Da) represents aligned peaks that are heterogeneous: they encompass peaks with different deduced accurate masses within the aligned mass bin, and thus represent potential cases of hyper-alignment. The majority of these cases are to be found in the densely populated 'washing phase' (P3) at the end of the chromatogram (figure 3). In the remainder of the chromatogram (P1 and P2), 167 of the 1079 aligned peaks (15.4%) are classified as heterogeneous (F), and thus potentially reflect (partially) misguidedly aligned mass peaks. However, the fraction of wrongly aligned mass peaks will be considerably lower, as the misalignment of a few mass peaks can already result in an increased SD of the accurate masses of the aligned peak, whilst most of the included mass peaks are properly aligned. To further characterize these two groups (T, F) their distribution in the mass domain is plotted (figure 4), showing that this phenomenon is more frequent at higher masses (>800 Da).

3.2.2. *Assessing hypo-alignment*

Hypo-alignment would cause some peaks to match with noise, instead of with the corresponding peak in the duplicate chromatogram. The frequency of hypo-alignment was estimated based on the aligned pairs for which at least one of both peaks had a detectable peak intensity of >10 × local noise. In the scatter plot of figure 5, these cases show up in the two satellite clouds at the base of the graph. Assuming all pairs of which the peak intensities differ more than 5-fold to represent hypo-alignment, this results in an occurrence of 918 cases of hypo-alignment on a total of 13,146 aligned peak pairs (7.0%). As this type of misalignment will result in the formation of two cases of hypo-aligned peak pairs, the actual number of mass peaks involved is 459 out of 12,687 (3.6%).

3.3. *Comparing metabolic profiles*

Two strategies for data analyses were employed to reveal the differences in mass peak profiles between the samples studied: (i) pair-wise comparisons between groups of samples representing e.g. two genotypes, or two harvest points, and (ii) the overall comparison of all samples in a single multivariate analysis. The first type of analysis, a direct comparison between groups of samples, is facilitated by MetAlign which includes the

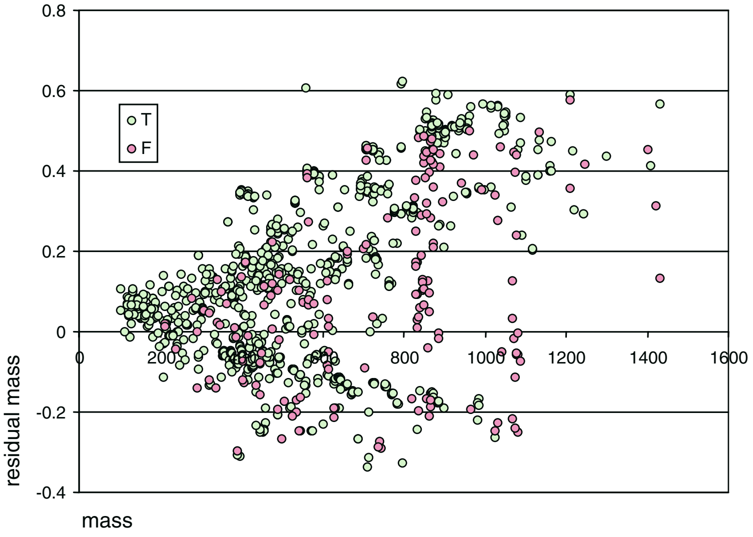


Figure 4. Residual masses (= average (accurate mass) − bin mass) plotted vs. mass. Same dataset as for figure 3, but limited to aligned peaks from 'flow-through' (P1) and 'gradient' phase (P2) of the chromatogram (figure 1a). Aligned peaks with SD < 0.029 are designated as T, the remainder as F.

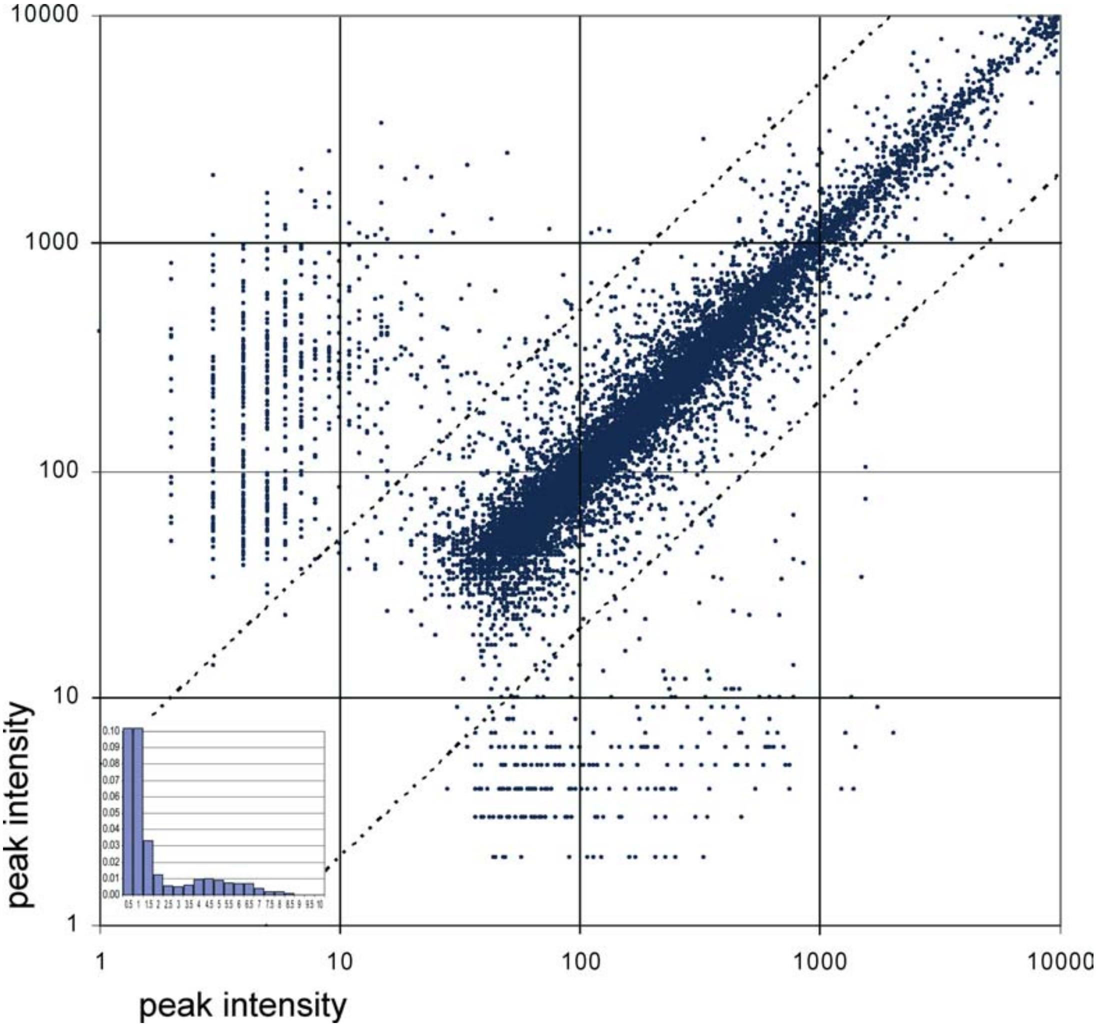


Figure 5. Scatter plot of the peak intensities of the aligned peak pairs of all technical repeats. The dotted lines indicate a 5-fold difference. The insert shows the distribution of the intensity ratios on a log scale.

option to perform a student *t*-test based extraction of significantly differential mass peaks at user-defined thresholds of probability, signal/noise ratio of peaks to include and the extent of change. The result of such a metAlign analysis, in which two samples from a single potato genotype (674) harvested at two different time points (e, l) have been compared in duplicate (1, 2), is shown in figure 6. Out of the total of 4106 aligned mass peaks, 102 peaks (figure 6c) are significantly ($p < 0.01$) at least 3-fold up-regulated in the early harvest samples (e1, e2), while only 10 peaks (figure 6d) have at least a three times higher abundance in the mature tuber samples (l1, l2; table 1). An analysis of the two other possible combinations of these four samples ([e1, l1] vs. [e2, l2] and [e1, l2] vs. [e2, l1]) reveals only zero and three significant differences, respectively, thereby demonstrating the significance of the revealed differences in the truly contrasting samples. It should be stressed that this total of 112 differential mass peaks certainly represents a lower number of metabolites, since isotope peaks, adducts and fragment ions will be included as well. However, through the accurate retention alignment, in combination with mass accuracy, at least the isotopes can easily be recognized from their similar retention times (or scan number). For instance, at scan 1213 the aligned masses 303.0504 and 304.0556 correspond to $C_{15}H_{11}O_7$ ($[M+H]^+$) and its C^{13}-isotope, respectively. The assumption that all differential peaks detected in a window of three scans are derived from a single compound would give an estimated number of 44 up-regulated compounds (35 in phase P2) and 8 down-regulated ones (4 in P2).

In the second strategy, involving a multivariate analysis of all samples studied, metAlign was used to generate a large data matrix (aligned peaks × samples) in which aligned peaks (in the rows) are identified by an accurate retention time [scan number] and mass bin [1 Da units]. Only peaks reaching at least 10× local noise were retained for further analysis. Figure 7 shows the results of a PCA on these data. First, it is evident that the duplicate sample extractions, included to give an indication of the technical variation, are close together

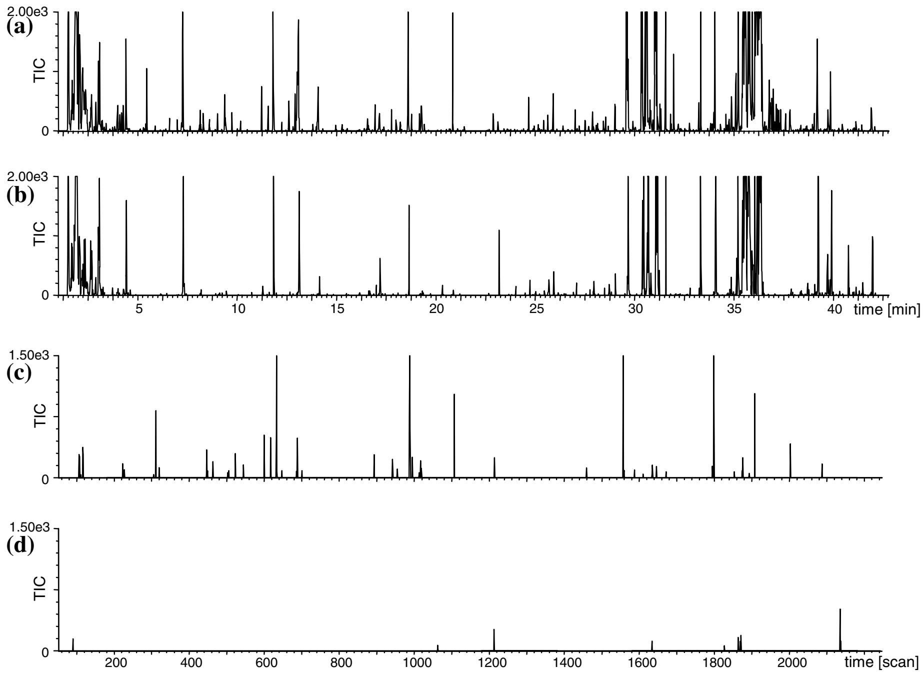


Figure 6. Differential analysis of two groups of samples using metAlign software. The groups correspond to the early and late harvested tubers of the same potato genotype (674) and have both two members (replicates). Shown are the calculated peak chromatogram of an early (a) and a late harvest (b) sample. Peaks significantly different between both groups ($p < 0.01$) are shown when at least 3-fold up in either early (c) or late (d) tubers.

as compared to the distances between most of the different biological samples (i.e. biological variation). Second, the samples cluster together into two distinct groups that correspond to the time of harvest. The centroids of both clusters clearly point to PC2, plotted on the *y*-axis, as being the component containing the harvest time-associated variation. PC1 can be seen as representing harvest time independent variation between the individual genotypes (i.e. genetic variation).

3.4. *Differential mass peaks*

To investigate further which compounds most contribute to the observed segregation of the metabolic profiles in the two harvest times, 20 aligned peaks that had highest PC2 values were selected, as indicated in the corresponding peak plot (loadings; figure 7b). In a hierarchical cluster analysis on the profiles of all 1818 aligned mass peaks, these peaks end up close together in a minimal cluster with 38 members. This harvest-time-specific cluster is shown in figure 8. Based on the strong clustering of mass peaks aligned at the same scan number, not only isotopes but also unavoidable in-source fragmentation could be identified. This result reflects the precision of the experimental and computational procedures. For instance, of scan number 590 both *m*/*z* 822.3054 and its C^{13}-isotope are clustered, while of scan 974 *m*/*z* 695.3624 as the most likely parent (i.e. the most abundant peak), some of its in-source fragments (*m*/*z* 531.3202 and *m*/*z* 123.0451; see below), as well as isotopes derived thereof are clustered (figure 8).

One of the more abundant peaks that was clearly specific to the early harvest time ($p = 0.000021$), having an accurate mass of 695.3624 Da ($[M + H]^+$), was selected for further analysis. This mass peak was subjected to accurate mass MS/MS in an attempt to elucidate its chemical identity (figure 9). The isotopic pattern of this single charged ion revealed the absence of sulphur, while its mass pointed to an even number of nitrogen atoms. Within 5.0 ppm mass accuracy, the exact mass measured revealed 8 possible elemental structures. Collision-induced fragmentation resulted in fragments with accurate masses ($[M + H]^+$) corresponding to $C_9H_9O_3$ (e.g. hydroxy-cinnamic acid), $C_7H_{16}N_3O_5$, $C_{16}H_{25}N_2O_3$, losses of NH_3, and combinations thereof. These fragments all point to a parent of $C_{32}H_{51}N_6O_{11}$ ($[M + H]^+$). To our knowledge a compound having this elemental composition in potato is yet unknown. Compounds in the SciFinder (Chemical Abstracts Service) database fitting the obtained elemental composition do, however, not comply with the observed MS/MS fragmentation pattern. The QTOF MS/MS-fragments alone did not allow for an unequivocal identification of this novel compound and additional ion-trap MS and NMR analyses are thus needed to elucidate its chemical nature.

Interpretation of many LC–MS chromatographic spectra in a single analysis is a major challenge, especially when the studied samples represent a certain amount of heterogeneity. Here we present a novel and unbiased strategy for extracting relevant information from complex multiple full-scan LC–MS-derived meta-

Table 1
Mass signals of the aligned peaks significantly different between early (674a) and late (674b) harvested tubers of potato genotype 674 based on duplicate analyses ($p < 0.01$)

2log(ratio)	Scan	*m*/*z*	674a-1	674a-2	674b-1	674b-2	*p*
7.20	2135	385.2054	–	–	524	510	0.0001
5.47	90	488.7930	–	–	158	148	0.0002
5.36	2137	386.2091	–	–	127	125	0.0005
5.26	1864	(1442.9813)	–	–	156	185	0.0011
4.78	1871	(1439.1580)	–	–	207	191	0.0019
4.53	1869	(1444.0605)	–	–	122	127	0.0042
4.40	1827	(1030.7577)	–	–	64	73	0.0056
2.45	1063	467.1196	18	13	92	82	0.0077
1.98	1213	304.0556	15	16	67	58	0.0094
1.68	1213	303.0504	95	100	336	294	0.0099
−6.73	435	177.0565	293	294	–	–	0.0000
−6.68	933	922.3362	216	240	–	–	0.0028
−6.55	286	531.3216	771	897	12	–	0.0057
−5.93	511	792.2915	231	225	–	–	0.0001
−5.87	215	163.0414	179	173	–	–	0.0003
−5.84	451	603.2650	188	213	–	–	0.0039
−5.74	1019	693.3537	110	127	–	–	0.0048
−5.72	976	533.3259	151	161	–	–	0.0012
−5.33	1594	(873.1329)	170	152	–	–	0.0035
−5.32	1553	1047.5393	4121	4068	115	90	0.0001
−5.15	434	265.1567	88	87	–	–	0.0001
−5.12	606	777.3035	144	146	–	–	0.0001
−5.09	215	251.1421	94	110	3	–	0.0067
−5.05	643	177.0568	98	85	–	–	0.0053
−4.91	1200	866.4908	245	266	–	–	0.0017
−4.90	976	–	97	111	–	–	0.0046
−4.86	976	693.3528	1344	1585	61	39	0.0072
−4.79	985	1130.3922	164	179	–	6	0.0020
−4.77	1571	850.4930	95	108	–	–	0.0039
−4.75	945	952.3483	109	114	–	–	0.0007
−4.72	985	1131.3933	88	95	–	3	0.0018
−4.70	686	736.6658	78	89	–	–	0.0047
−4.66	976	–	103	116	–	–	0.0039
−4.66	283	367.2738	119	132	–	–	0.0028
−4.63	977	744.3537	106	112	–	–	0.0009
−4.63	620	489.2466	129	141	–	–	0.0021
−4.60	976	532.3250	431	489	31	–	0.0051
−4.57	1555	1045.5130	90	100	–	–	0.0030
−4.55	689	472.2472	87	94	–	–	0.0015

Of the down-regulated peaks, only the 29 most differential ones are shown. Accurate masses (*m*/*z*) are averages from all aligned peaks (30) analyzed; brackets indicate unreliable masses as they represent potential cases of hyper-alignment.

bolic profiles, using metAlign software to produce a matrix of aligned mass signals (intensities of aligned mass peaks × samples) and subsequent statistical analyses tools for unsupervised data analyses. As a proof of principle, crude aqueous methanol potato tuber extracts were analyzed by HPLC-QTOF MS and complete LC–MS profiles, including the chromatographically problematical injection peak and column washing regions, were used for processing and alignment. PCA of the aligned mass signals revealed trends that were in concordance with the underlying structure of the data, thus demonstrating the overall validity of the performed experimental and computational procedures as a whole. Moreover, we show that this non-directed approach is very successful in detecting differential mass signals, even in the case of a dataset derived from heterogeneous biological samples. The strength of using metAlign for this purpose is its capability of combining automated background subtraction based on local noise calculation, extraction of mass peak information, and sophisticated alignment of the chromatographic axis. The baseline corrected data, as calculated by metAlign, can be visualized in the original LC–MS software in order to evaluate the extent and success of baseline subtraction and mass trace mining (e.g. figure 1). The proper alignment of gradient-based HPLC chromatograms cannot be performed by a simple retention time correction, because under these conditions retention time shifts can occur even within a single run, e.g. due to small changes in temperature, acidity and compound

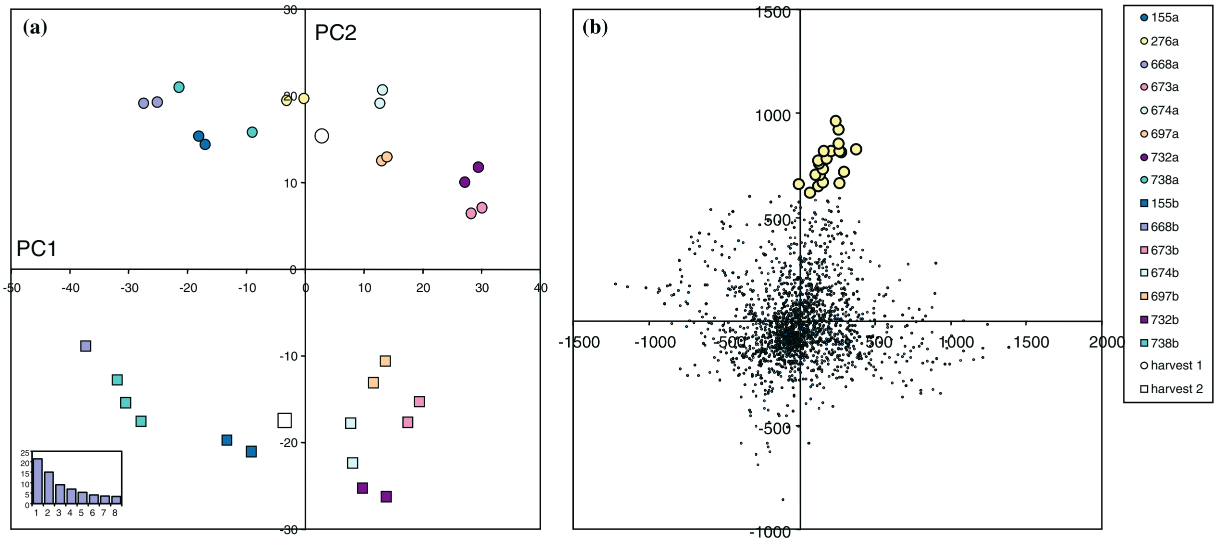


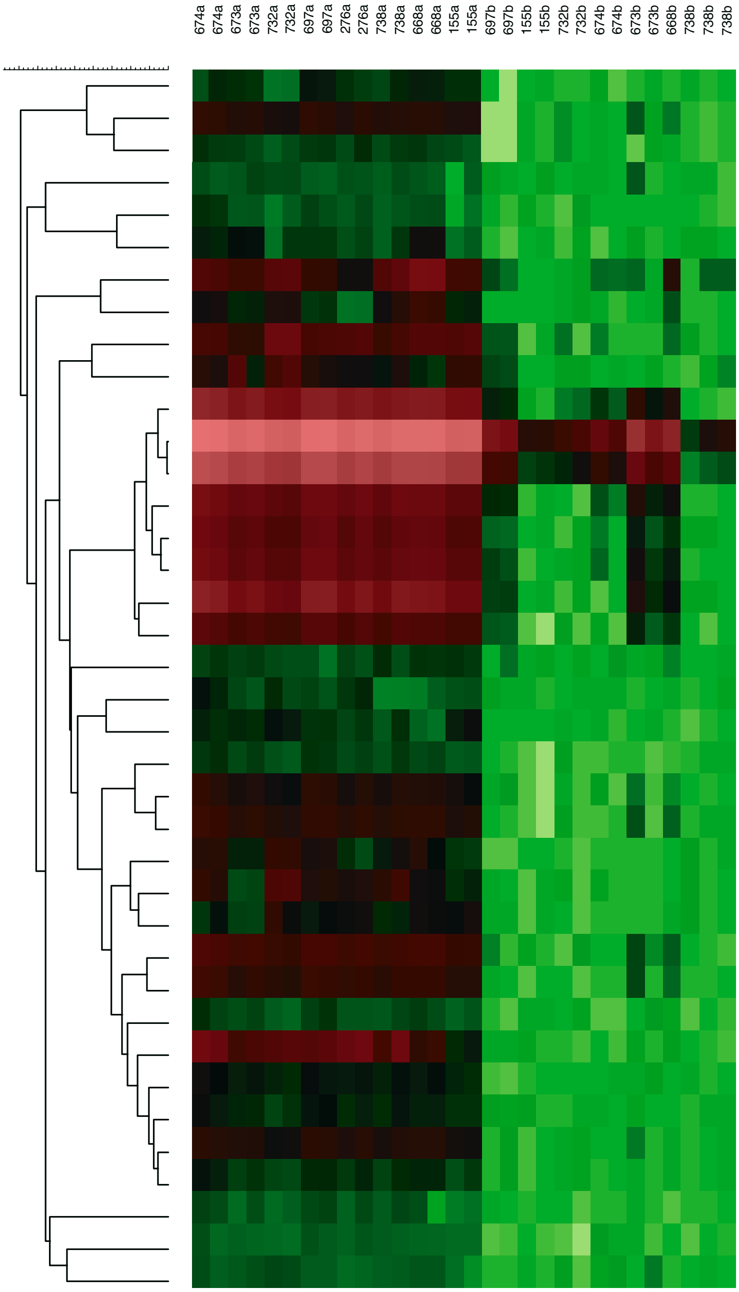
Figure 7. (a) Principal component analysis (PCA) of metabolite profiles of tubers from eight potato genotypes at two harvest times. Colours indicate the different potato genotypes, symbol shape corresponds to harvest time (circles: early harvest; squares: late harvest). The larger open circle and square are the centroids of the early harvested and late harvested tubers samples, respectively. The insert shows the percentage of variance explained for each of the first eight PCs. (b) Corresponding plot of the analyzed mass peaks. The 20 aligned peaks highlighted in yellow were selected for further analysis.

concentration. MetAlign utilizes non-linear algorithms based on the local retention shifts of land-mark peaks (i.e. common mass traces) present in the chromatograms. Without such accurate background correction and non-linear alignment, it is unfeasible to make a detailed comparison of a large number of LC–MS analyses in a non-directed manner.

After alignment, a pair-wise comparison of groups of samples can be made within metAlign, in order to identify all significant differences between the mass profiles of the groups and differential chromatograms can be generated based on the absolute or the relative differences in mass signals. Alignment of many datasets derived from biologically more heterogeneous samples is a more difficult task than just comparing analyses derived from only two samples. Here we show, using an example of 32 analyses of eight different potato genotypes at two harvest times and in duplicate, that by the use of metAlign it is possible to combine such large datasets and that subsequent PCA (figure 7) helps to uncover the underlying structure of the data: replicates being highly similar, while the second principal component (PC2) clearly associates with the harvest time and PC1 corresponds with harvest-time independent differences between the genotypes.

The potential weak point of this metAlign-based approach is its restriction to retention time and molar mass as the basis for combing mass peaks. As a result mass peaks that represent distinct compounds may be combined (hyper-alignment), while on the other hand cognate mass peaks are not recognized when the algorithm is too reluctant in aligning peaks (hypo-alignment). The extent of hyper-alignment was evaluated on the multiple aligned dataset. Especially the 'washing phase' (P3) of the chromatogram is prone to hyper-alignment due to the high density of peaks and the abundance of multiple-charged ions (figure 3). Outside this region, 15.4% of the aligned peaks is heterogeneous (as based on SD of deduced accurate masses). This figure might in part reflect errors in the deduced mass calculations, but could also indicate the combination of distinct mass peaks. Several reasons could play a role, e.g. the presence of more than one peak in a single 1 Da mass bin, either representing distinct compounds or multiple charged ions derived from the same compound. Another type of explanation might be sought in cases were the residual mass size is close to the applied limits of the mass bin (from −0.35 to +0.65 Da), in which case identical peaks might end up in neighbouring bins. In our samples this effect gained importance in masses above 800 Da where the residual mass starts to approach the limits of the bin at this value (figure 4). By simply decreasing the mass bin size, the algorithm would gain from the accurate mass information to reduce the incidence of the former phenomenon, but would negatively effect the latter by an increased 'border effect'. Therefore mitigation of this type of problem must come from carefully tailored chromatography as well, thus preventing the occurrence of such regions in the chromatogram by an enhanced resolution (e.g. Tolstikov *et al.*, 2003).

Hypo-alignments played a role in 3.6% of the mass peaks. As duplicate runs were done on replicate extractions of the same material, these inconsistencies partly reflect analytical variation. However, some of



	nr	P-value	m/z
	661	0.0044	807.3133
●	975	0.0001	743.3521
	974	0.0000	744.3537
	687	-	337.0969
	974	0.0000	533.3259
	1017	0.0636	693.3537
●	590	0.0514	822.3054
	590	0.0305	823.3084
●	608	0.0118	776.2982
	1626	0.1619	(832.1054)
●	974	0.0001	531.3202
●	974	0.0000	695.3624
●	974	0.0001	696.3690
●	975	0.0001	727.3578
●	974	0.0002	694.3558
●	974	0.0001	697.3793
●	974	0.0004	693.3528
●	975	0.0001	709.3467
	77	0.0283	427.9152
	1875	0.0204	575.3617
	1627	0.0094	(827.4166)
	974	0.0001	712.3617
●	975	0.0001	710.3510
●	975	0.0001	711.3595
●	536	0.0044	792.2959
●	476	0.0311	776.2995
●	608	0.0060	777.3035
●	974	0.0001	532.3250
●	975	0.0001	728.3610
	975	0.0002	123.0451
●	301	0.0117	531.3216
	976	0.0000	761.3615
	975	0.0001	725.3403
●	975	0.0000	698.3774
	974	0.0001	(699.2144)
	974	0.0000	729.3579
	976	0.0001	653.3550
	680	0.0040	736.1614

Figure 8. Cluster containing the 20 early harvest associated mass peaks (high PC2 values) after cluster analysis on all 1818 aligned peaks, based on Pearson correlation of the log-transformed amplitudes. The peaks labelled with yellow dots are those highlighted in figure 7b. For each aligned peak, scan number (nr), *p*-value (early vs. late harvested tubers) and calculated accurate mass (*m*/*z*) are indicated. Accurate masses are averages from aligned *m*/*z* signals of all samples analyzed; brackets indicate unreliable masses, as they are derived from potential cases of hyper-alignment.

them will be caused by the still imperfect nature of the computational procedure applied.

4. Concluding remarks

In conclusion, this work shows that the use of software specifically designed for non-targeted metabolomics applications helps to discover differences in large LC–MS derived datasets in the absence of any prior knowledge. Once it is known which mass peaks are specifically different, subsequent identification efforts can then be concentrated on the relevant compound(s). This approach represents a significant step forward when dealing with multiple complex chromatographic spectra but of course there are still some limitations which require attention.

It can be argued, that the more dissimilar the studied samples are, the more difficult it is to obtain a perfect matching of compounds based on retention time and molecular mass information only. The present study

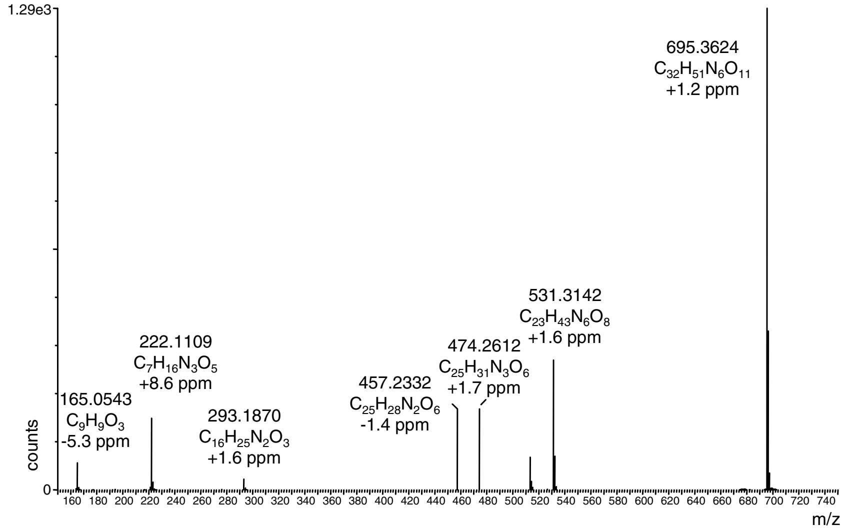


Figure 9. On-line MS/MS of an early-harvest specific ion. Fragmentation was performed by a step-wise increase in collision energy (5, 15, 30 and 60 eV) on mass 695,3624. Note that mass signals between *m/z* 440 and 560 are magnified by a factor 36.

on eight different potato genotypes at two different physiological states (harvest time) has shown that it is possible to uncover differential compounds in a moderately heterogeneous context. This opens up possibilities for the untargeted analysis of the complex metabolic profiles of large collections of samples, e.g. mutant libraries, breeding populations, and natural populations.

Acknowledgments

We would like to thank Chris Maliepaard for valuable discussion. The presented research was carried out within the research programme of the Centre of BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

References

Aharoni, A. and Vorst, O. (2002). DNA microarrays for functional plant genomics. *Plant Mol. Biol.* **48**, 99–118.

Bino, R.J., Hall, R.D., Fiehn, O., *et al.* (2004). Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* **9**, 418–425.

Cánovas, F.M., Dumas-Gaudot, E., Recorbet, G., Jorrin, J., Mock, H.P. and Rossignol, M (2004). Plant proteome analysis. *Proteomics* **4**, 285–298.

Celis Gamboa, B.C. (2002) *The life cycle of the potato*(Solanum tuberosum *L.*)*: from physiology to genetics*. Ph.D. thesis, Wageningen, 186 pp.

Fernie, A.R., Trethewey, R.N., Krotzky, A.J. and Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* **5**, 763–769.

Fiehn, O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171.

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

Gallardo, K., Job, C., Groot, S.P., *et al.* (2002). Proteomics of *Arabidopsis* seed germination. A comparative study of wild-type and gibberellin-deficient seeds. *Plant Physiol.* **129**, 823–837.

Goossens, A., Hakkinen, S.T., Laakso, I., *et al.* (2003). A functional genomics approach toward the understanding of secondary metabolism in plant cells. *Proc. Natl Acad. Sci. USA* **100**, 8595–8600.

Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L. and Bino, R. (2002). Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell* **14**, 1437–1440.

Hirai, M.Y., Yano, M., Goodenowe, D.B., *et al.* (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **101**, 10205–10210.

Lommen, A., Weseman, J.M., Smith, G.O. and Noteborn, H.P.J.M (1998). On the detection of environmental effects on complex matrices combining off-line liquid chromatography and ^{1}H-NMR. *Biodegradation* **9**, 513–525.

Meyers, B.C., Galbraith, D.W., Nelson, T. and Agrawal, V. (2004). Methods for transcriptional profiling in plants. Be fruitful and replicate. *Plant Physiol.* **135**, 637–652.

Noteborn, H.P.J.M., Lommen, A., Jagt, R.C.van der and Weseman, J.M. (2000). Chemical fingerprinting for the evaluation of unintended secondary metabolic changes in transgenic food crops. *J. Biotechnol.* **77**, 103–114.

Roessner, U., Luedemann, A., Brust, D., *et al.* (2001). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**, 11–29.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000). Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* **23**, 131–142.

Roessner, U., Willmitzer, L. and Fernie, A.R. (2001). High-resolution metabolomic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* **127**, 749–764.

Schnable, P.S., Hochholdinger, F. and Nakazono, M. (2004). Global expression profiling applied to plant development. *Curr. Opin. Plant Biol.* **7**, 50–56.

Soga, T., Ueno, Y., Naraoka, H., Matsuda, K., Tomita, M. and Nishioka, T. (2002). Pressure-assisted capillary electrophoresis electrospray ionization mass spectrometry for analysis of multivalent anions. *Anal. Chem.* **74**, 6224–6229.

Sumner, L.W., Mendes, P. and Dixon, R.A. (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817–836.

Tolstikov, V.V. and Fiehn, O. (2002). Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.* **301**, 298–307.

Tolstikov, V.V., Lommen, A., Nakanishi, K., Tanaka, N. and Fiehn, O. (2003). Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem.* **75**, 6737–6740.

Von Roepenack-Lahaye, E., Degenkolb, T., Zerjeski, M., *et al.* (2004). Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* **134**, 548–559.

Wagner, C., Sefkow, M. and Kopka, J. (2003). Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* **62**, 887–900.

Watson, B.S., Asirvatham, V.S., Wang, L. and Sumner, L.W. (2003). Mapping the proteome of barrel medic (*Medicago truncatula*). *Plant Physiol.* **131**, 1104–1123.

Weckwerth, W. (2003). Metabolomics in systems biology. *Annu. Rev. Plant Biol.* **54**, 669–689.

Weckwerth, W., Wenzel, K. and Fiehn, O. (2004). Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* **4**, 78–83.

Wolff, J.-C., Eckers, C., Sage, A.B., Giles, K. and Bateman, R. (2001). Accurate mass liquid chromatography/mass spectrometry on quadrupole orthogonal acceleration time-of-flight mass analyzers using switching between separate sample and reference sprays. 2. Applications using the dual-electrospray ion source. *Anal. Chem.* **73**, 2605–2612.