Metabolite Peak Identification and Data Structure in a Multi-Site, Large Scale Metabolomics Experiment

John Draper*[1], Manfred Beckmann[1], Scott Campbell [2]; Derek Stewart,[3] Wynne Griffith,[3] Rhoda Marshall, [3] Susan Verrall [3]

* e-mail jhd@aber.ac.uk

[1]Department of Biological Sciences, University of Wales Aberystwyth, Ceredigion, SY23 3DA, UK;
[2] SpectralWorks Ltd, The Heath Business & Technology Park, Runcorn, WA7 4QF, UK;
[3] Scottish Crop Research Institute, Quality Health and Nutrition, Invergowrie, Dundee, DD2 5AJ

DS, WG and SV sponsored by SEERAD

Introduction

If metabolomics approaches are to have true utility in Functional Genomics programmes it will be important to tackle several important areas:

- to measure simultaneously as many metabolites as possible to approach ëglobalícoverage within the constraints imposed by biological and analytical variance
- to develop sufficient replicate chromatograms that can be compared meaningfully in order to have sufficient data to utilise supervised data analysis techniques.
- to develop a standardised data model that can be used to capture data from a range of instrument platforms in different laboratories and develop an international data-base strategy
- to have an overall experimental statistical design that will allow meaningful comparisons to be made of many hundreds of genotypes.

Currently, the majority of reported GC-MS profiling experiments have analysed less than 100 chromatograms and made comparison of only a few (4-12) genotypes and in some instances have measured the relative ratios of only around 100 abundant metabolites for which standards exist. Lack of standardised methods for the pre-processing of raw GC-MS data is a major factor inhibiting development of an international database strategy for metabolomics. Although ëringtestingí approaches are common to validate quantitative, targeted analysis of a few compounds in complex biological samples, to date there are no examples published of co-operative metabolomics experiments which require the high throughput measurement of the relative ratios of many hundreds of metabolites in the same extracts in more than one laboratory. The present work seeks to identify some of the major bottlenecks in order to develop a future strategy for metabolomics.

Results & Discussion

(1) Peak finding and spectrum deconvolution in metabolome profiling

Sensitive GC-tof-MS technology with high scan rates allows theoretically the user to resolve in excess of 1000 metabolite peaks in a single run. In crude extracts the first challenge is to reproducibly identify peaks representing sample chemistry and deconvolve spectra of often >600 metabolite peaks present at a large dynamic concentration range in each chromatogram. Using instrument manufacturers' software combined with manual checking of peak alignment we have processed data representing 2304 very similar samples. In general **more than 70% of peaks represented 'unknown' metabolites** with no, or very poor, matches in spectral databases. Such peaks are rarely annotated in a meaningful way for any subsequent multivariate analysis, which is exacerbated in inter-laboratory data mining experiments. Particularly evident is the problem relating to the generation of false positive peaks (**Figure 1**) in which aberrant peaks (**false positives**) are identified by instrument software often coincides with the loss of a previously validated peak such as leucine in the example shown.

The manual pre-processing of this large amount of metabolomics data requires skilled individuals and takes longer than the chromatography runs themselves, thus soon becoming a major bottleneck. However, the end result is that detailed comparisons can be made of the metabolic profiles of closely related samples to generate new insight into genotype related differences (**Figure 2**). **The overall aim of the present study is to try and automate data pre-processing to remove this bottleneck and to start to develop approaches by which data produced on different machines can be compared meaningfully.**

Figure 1. False positive and false negative peaks in GC-MS peak tables

Although expected in all samples leucine is deconvolved in only 1 of 3 replicated runs in set C1. In two further sets leucine is not deconvolved in any of the runs and instead the instrument software automatically deconvolved extraneous peaks from system noise at the retention time of leucine

Run	Peak	R.T. (sec)	Area	Name	Spectra
C1	1	251.841	29894	Leucine di-TMS	**158:1998** 147:424 102:309 133:289 159:284 116:252 100:195 103:165 211:139 115:105
	2	252.541	40926	Unknown	**138:1223** 108:433 182:296 197:221 85:180 80:156 139:148 84:147 83:141 158:97 (..)
	3	252.891	64312	Tungsten, pentacarbonyl(..)	**93:1872** 147:949 **98:565** 95:449 94:232 84:181 80:165 131:127 148:127 96:113 (..)
C2	1	251.575	39452	Tungsten, pentacarbonyl(..)	**158:1435** 93:970 147:681 95:381 116:354 102:334 159:235 100:215 84:211 103:205 (..)
	2	252.275	30816	Tungsten, pentacarbonyl(..)	93:797 **138:735** 147:656 95:324 108:322 84:316 85:316 139:235 182:223 83:203 (..)
	3	252.875	23102	2, 4a-Epoxymethylphenanthrene(..)	98:1137 147:451 93:450 169:137 95:119 99:108 113:91 101:84 96:78 111:78 (..)
C3	1	251.225	8590	Unknown	**93:621** 127:288 94:229 80:227 115:215 83:158 472:152 97:149 130:148 **138:80** (..)
	2	**missing**			m/z 138 assigned to peak 1;
	3	251.775	7719	Tungsten, dicarbonyl-(..)	147:1014 **98:374** 84:329 85:298 117:216 113:206 100:201 148:195 103:157 101:150 (..)

(2) False positive peak accumulate in large scale metabolomics experiments

As the vast majority of peaks identified in metabolome profiling are unknown there are no standard spectra available for comparison. Unknown peaks are generally also of relatively low intensity and so deconvolution of spectra from background noise, co-eluting and flanking peaks is also difficult. Under these circumstances peak annotation becomes a major problem as high confidence matches with previously identified peaks in a user library are not always achieved. The end result is the gradual accumulation of false positive peaks as more and more chromatographic runs of similar extracts are processed.

The Pegasus II GC-tof-MS instrument manufactured by LECO Corp. is one of very few instruments with bespoke software which allows automated peak finding and spectrum deconvolution. In the example shown in **Figure 3** the peak finding parameters in the LECO system were set to identify 1000 unknown peaks in a GC-MS chromatogram of a crude plant extract using automated peak table generation. Allowed to operate automatically without user intervention after 72 runs the software had already defined presence of over 3000 ëpeaksí in a matrix that probably contained less than 1000 independent metabolites.

Figure 2. PCA-DFA of potato cultivars by analysis of GC-MS data

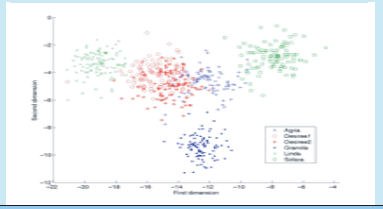


Figure 3. Accumulation of false positive peaks in large metabolomics experiments

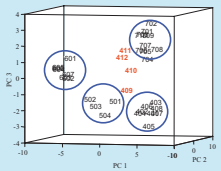
Preliminary peak alignment of X runs on LECO system with on average 500 deconvolved peaks results in Y common peaks:

X	Y
8	700
16	1300
29	2600
72	3500

(3) Machine drift makes it difficult to compare peak tables generated over a long time period in metabolomics experiments

In metabolomics GC-MS profiling experiments metabolites are not quantified against standards but instead normalised within a run to some chromatogram parameter such as total peak area or internal standards and data presented as relative ratios. Run quality is assessed by examining the chromatographic behaviour of a quality control mixture of 20-30 standard chemicals to monitor peak retention time shifts and intensity change thus allowing runs to be rejected that fall outside pre-determined thresholds. However, column aging and periodic maintenance cause both gradual drift and more abrupt changes in instrument responses. This results in time batch related clustering of samples when analysed by multivariate data analysis techniques such as Principal Components Analysis (PCA). For example **Figure 4** illustrates time batch clustering of GC-MS QC standards over a period of 4-5 months. Such behaviour is extremely difficult to calibrate between runs and confounds facile data analysis, particularly by unsupervised methods.

Figure 4 Monitoring GC-MS ëbatch reproducibilityí based on PCA of response of Quality Control Mixtures

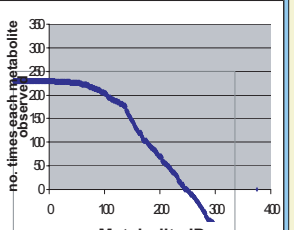


Number groupings refer to major events such as ion source cleaning, system tuning, column cutting and column changes over 4-5 month period

(4) A combination of false positives, fluctuations in individual metabolite concentrations between samples and instrument sensitivity complicates comparison of replicate chromatograms

Large-scale metabolomic experiments require analysis of many hundreds of samples over long periods of time. To compensate for instrument variance it is extremely important to randomise injection order of samples in order to generate data that is not skewed by instrument drift. Metabolites are present in plant tissue extracts in a dynamic concentration range, which can differ over several orders of magnitude between individual metabolites. Thus many low-concentration metabolites will regularly fall below the threshold for accurate detection in many runs, depending on system sensitivity at the time and absolute concentration in a replicate sample. A zero value for a metabolite occurs because a peak is not found by the instrument (**Figure 5**), either because the concentration is too low (missing value) or because a false positive peak has been generated in the same retention time window and thus the expected peak is not found (false negative). In the case of missing values due to concentration differences between metabolites there is a linear relationship between the number of zero values for a metabolite in a series of data tables and its expected signal intensity (**Figure 6**). Unlike the normal expected behaviour of peaks, a possible false positive peak among the lower concentration metabolites can often be identified because they will be shifted to the right of the expected curve due to the fact when actually measured their intensity will be higher than predicted by their frequency of zero values in the data table.

Figure 5. NO. of times metabolites found in LECO GC-tof-MS data set containing 2340 samples and 370 expected peaks



(5) Comparison of Peak Tables generated following analysis of same extract on three different GC-MS machines

A major step in the integration of metabolomics data in the future is the ability to analyse similar extracts on different instruments based in different laboratories and to generate data tables in which all the peaks are aligned and may thus be utilised in meaningful data mining experiments. In a preliminary study to assess the scope of this challenge we compared chromatograms generated on three instruments using a common extract. Polar extracts of potato tubers were prepared using water, methanol and chloroform, dried down and derivatized by substituting labile hydrogen atoms with a trimethylsilyl-group. Samples were analysed on three different GC-MS instruments using columns with similar polarity (e.g. DB5), but different injector/detector technologies, run parameters and scan rates.

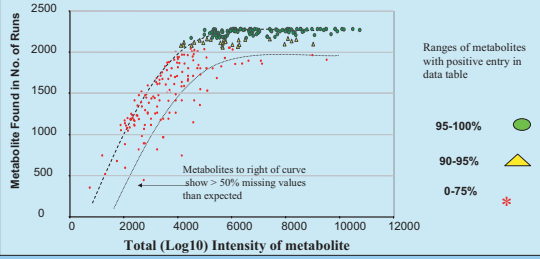
System: Agilent MSD Quadrupole GC-MS, Finnigan Tempus GC-tof-MS, LECO Pegasus III GC-tof-MS

***Solanum* samples** Desiree, Phureja, Cara

Manual comparison of GC-MS metabolite peak lists from the instruments indicates perfect correspondence of **elution order** of known metabolites. The majority of **major known metabolites** (approx. 50 peaks, 20 % of total number found using all instruments) are found in all three peak lists. TOF-systems detect an additional 20-25% known metabolite peaks reproducibly found in replicated runs. Many of the higher intensity unknown peaks (approx. 50, 20 %) recognisable by spectra and RT correspond between instruments. A large percentage (approx. 40 %) of lower intensity unknown peaks do not correspond between peak tables generated on different machines. It is not certain if these represent artefactual peaks in some instruments, but currently they are identified reproducibly in at least one system.

For a multi-site metabolomics project to function it is essential that all variables (metabolite peaks) in data analysis can be aligned and as the vast majority of peaks are low abundance and unknown then a rational procedure has to be put in place to correct or remove such peaks from data tables.

Figure 6. Expected relationship between metabolite signal intensity and frequency of missing values in GC-tof-MS data table



(6) Developing approaches to deal with peak alignment in output from different GC-MS instruments using metabolome profiling data

The presence of large numbers of zero values for some metabolites in a data table results in non-normal and disjunctive distribution of peaks in replicate data tables which confounds many types of commonly used unsupervised statistical analysis techniques such as ANOVA and PCA. This is a major problem in metabolomics and particularly the presence of false positive peaks is something that should be possible to test for in future. Two possible related approaches are being researched to try and improve data quality in GC-MS profiling. As a longer term solution it should be possible to develop more robust peak finding (and peak spectrum deconvolution) algorithms to analyse raw data from any instrument that will require any new ëfoundí peaks to conform to a greater range of ëexpected behaviouríí **between many replicate runs** and not just within a single run. Thus new peaks would not be allowed to occupy a user library of expected peaks until being validated. Secondly, as a shorter term solution it should be possible to develop software that will analyse data tables generated automatically by manufacturers software and highlight found peaks that may not conform to expected behaviour which can then be left out of any data analysis.

As a first step in this process we have been using the software package AnalyzerPro (SpectralWorks Ltd) to assess its capacity to accept raw data files from the three GC-MS instruments and generate peak tables that can be compared for peak alignment. Initial experience is that AnalyzerPro accepts, processes, deconvolves and generates peak tables in a fully automated fashion of raw data, Net-CDF- and .CSV-files from all three systems. Running AnalyzerPro we were able to quickly process **automatically** raw data files representing different potato extracts and deconvolve between 100-600 peaks from each run. An example of the type of output obtained from this package is illustrated in **Figure 7.**

Manual pre-processing of data requires great expertise in mass spectrometry and uses up a substantial time resources. Our initial impressions are that peak alignment between replicate, consecutive runs made on the same instrument is excellent but comparison of runs made in different time frames is more problematic due to instrument drift. At this initial stage of the investigation somewhere between 30-35% of deconvolved peaks are aligned in output from all three instruments. Bearing in mind the difference in data acquisition parameters and sensitivity between the three systems this approach looks promising for further develop.

In the second approach we have started to assess the potential of the Matrix Analyzer (module in AnalyzerPro) to examine data table output from GC-MS instruments. Data from replicated runs of potato tuber extracts from each system were deconvolved and compared against a reference component list. The data matrix was prepared in Excel (**Figure 8)** for data analysis. Statistical analyses (PCA) of normalized peak areas (against internal standard) and generation of loading plots of matrices were performed in Matlab (The MathWorks, Inc). Component IDís were directly linked to components in Matrix Analyser. Peaks were annotated in loading plots if component matches retention index and mass spectrum of external standards analyzed under same conditions.
The PCA plots (**Figure 9**) indicate that potato genotypes were well discriminated using the data matrix generated by AnalyzerPro. Examination of the loadings plots identifies a small number of discriminatory metabolites which separate the species *Solanum phureja* from the two *Solanum tuberosum* varieties Desiree and Cara in PC1 (red). In PC2 Desiree (green) and Cara (blue) are well discriminated by individual metabolites.

In summary, our initial experience is that AnalyzerPro is a quick and useful tool to pre-process data tables for meaningful analysis and the process does work using data from all three types of GC-MS instrument. As expected the major problem associated with the whole process is the quality (reproducibility) of the initial deconvolution.

Figure 7. Example of data output following analysis of raw data files from a LECO GC-tof-MS by Analyzer Pro (SpectralWorks)

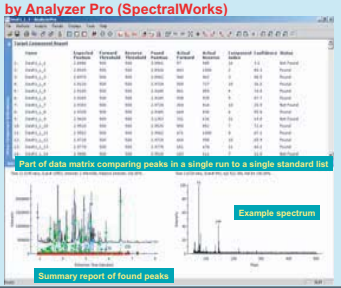


Figure 8. AnalyzerPro Data Matrix imported into Excel: alignment of GC-MS data in several runs

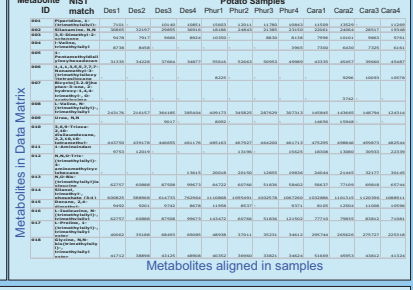


Figure 9. PCA of representative GC-MS data discriminating 3 potato genotypes