Universiteit
Leiden
The Netherlands

Quantification in untargeted mass spectrometry-based metabolomics

Kloet, F.M. van der

Citation

Kloet, F. M. van der. (2014, May 21). *Quantification in untargeted mass spectrometry-based metabolomics*. Retrieved from https://hdl.handle.net/1887/25808

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/25808

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/25808 holds various files of this Leiden University dissertation

Author: Kloet, Frans van der
Title: Quantification in untargeted mass spectrometry-based metabolomics
Issue Date: 2014-05-21

1

INTRODUCTION

In the Oxford Dictionary the metabolome is defined as: the **total number** of **metabolites** (the small molecules that are intermediates or products as a result of a metabolic reaction) present within an **organism**, **cell** or **tissue**. This definition covers the three key factors of metabolomics, the research field investigating the composition, role and function of the metabolome [140]. In the analysis of the definition of the metabolome we first identify the biological origin of the research field which can vary with regards to the type of biological question, and therewith connected, the type of samples that are analyzed. This can range from small individual cells to cells clusters to tissue slices to all sorts of biofluids like blood, urine or cerebrospinal fluid. Secondly, the term metabolite implies some form of identification of the chemical compounds being studied. Popular profiling and identification methods range from Nuclear Magnetic Resonance (NMR) to Mass Spectrometry (MS). Recently in particular fragmentation trees obtained with the MS^n [61, 106] approach are used to assign identities to the data features obtained with MS-based profiling techniques. The total number refers to the number of identified (and unidentified) chemical (metabolic) features and their concentration levels that are detected with the same analytical techniques mentioned before.

To obtain biological interpretable results these three important types of information (identity, quantity, and biological relevance) on a metabolite (Figure 1) are equally important and interact strongly. For a good understanding of the biological context the identity of the metabolites must be known. Conversely, identification of metabolites can be greatly improved by including biological information [58]. Furthermore proper (relative) quantification of the metabolites in question [139, 91] is necessary for a better understanding and modeling of the chemical processes of the biological system of interest, i.e. hypothesis generating metabolomics. Even though determination of the quantities of metabolites is not a necessity in all metabolomics experiments, it is very helpful for proper biological interpretation.

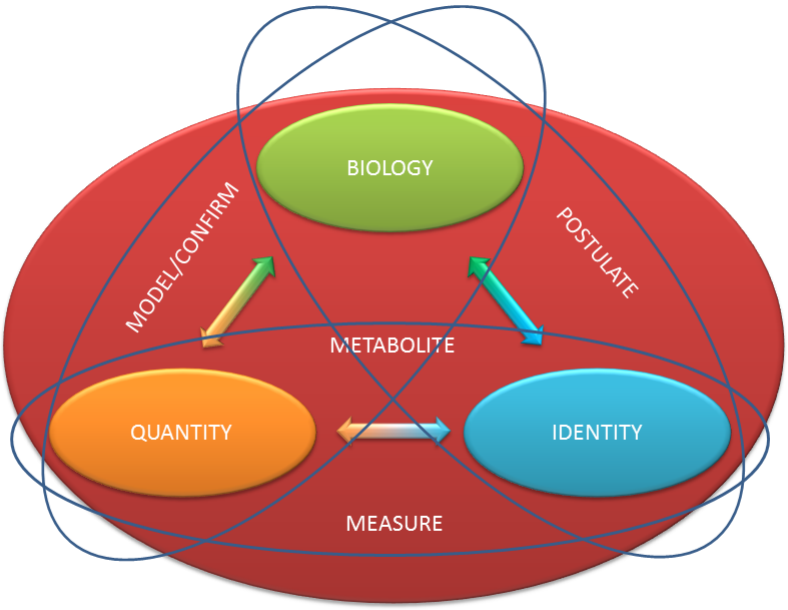

Figure 1: The three key factors of metabolomics: Biological relevance, Identity and Quantity and their interactions.

The word metabolome itself is a construct of the words metabolism and genome [47] and hints to the hierarchy within cell biology: the metabolome is the result of a whole range of chemical and regulation processes that are the result of the interaction of other biochemical organization levels such as the genome and their interaction with the environment. For example, changes in a cells physiological state as a result of gene deletion or overexpression are the complex result of processes at the transcriptome and the proteome and ultimately metabolome level[65, 126, 55]. For example, hard to detect multifactorial changes in the genome resulting in a disease may be easier detected by changes in the metabolite concentrations. This amplification of effects indicates the strength of metabolomics.

Contrary to proteomics and genomics the chemical structures and, therefore, physicochemical properties, observed within the metabolome are much more diverse. The proteome and genome consist of well-defined structural building blocks (i.e. amino acids and nucleotides respectively, although possible post-translational modification and epigenetics have to be taken into account). The diversity in the metabolome combined with the fact that metabolites are known to participate in many different biological pathways, reactions and processes challenges determination and biological interpretation in metabolomics. Without the proper biological knowledge there is no (bio-) logical explanation even if discriminating metabolites are found. The ubiquitous presence of fluids like for example blood at various places of possible biologically relevant processes, complicates the interpretation of metabolic activity in isolation even further[93] as they are not specific to any part of the body. To study a se-

lected part of cellular metabolic networks in a targeted manner, more recently tracer-based metabolomics has been developed as a new experimental data acquisition approach[77].

1.1 MASS SPECTROMETRY

Hyphenated mass-spectrometry (GC, CE or LC-MS) has become the predominant technology for determining metabolite abundances, mainly because of its sensitivity allowing the measurement of low abundant metabolites in small sample volumes. In Figure 2 the schematic of a time-of-flight (TOF) mass spectrometer (MS) detector is shown. The analytes are ionized in the ion source and separated by the applied electric field E (between grid A and B) in which the ions are differentially accelerated depending on their mass and charge. The time it takes to reach the detector (from B to C (length L)) is characteristic for the mass/charge ratio of an ion. Ions with a lower mass (having the same charge) are accelerated more and reach the detector earlier due to $E = \frac{1}{2}mv^2$ and consequently $m = \frac{2E}{v^2}$. v is the velocity i.e. the measured time (t) it takes for ions to travel the distance L.

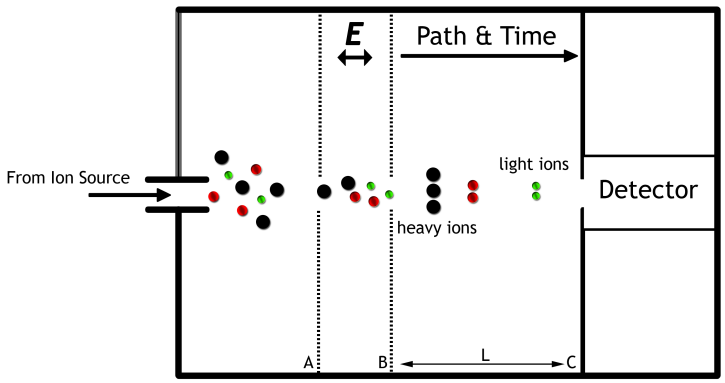

Figure 2: Schematic of a linear Time of Flight (TOF) mass spectrometer, heavier ions (with same charge) travel proportionally slower than lighter ions in an electric field.

When all ions have reached the detector a mass spectrum can be generated (Figure 3). The intensity on the y- axis corresponds to the number of ions that were detected with a specific mass (the x-axis).

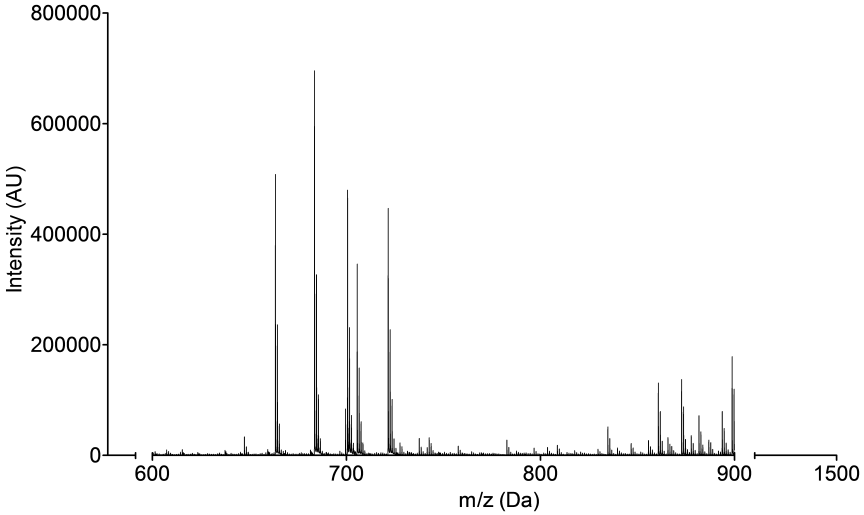

Figure 3: A typical (part of a) mass spectrum.

A drawback in MS is that each metabolite has its own response factor, i.e. the signal depends on the number of molecules but also on the type of molecule. For example two metabolites showing up in a mass spectrum with each of its (e.g.) protonated molecule having an intensity of 106 do not necessarily have the same concentration when they are introduced into the MS. This depends on factors like solubility, ionizability, fragmentation, etc. [4], which are different for the different metabolites. In addition, mass-dependent discrimination can occur due to the mass spectrometer. Furthermore, the response factor for a certain metabolite is matrix dependent, i.e. dependent on the composition of the solvent (in which various compounds can be present) when introduced into the MS, and consequently can vary over different samples creating differences in measured responses for identical metabolite concentrations[6, 79]. With other words, in two different human plasma samples the same metabolite with the same concentration can have different responses. The complex interactions between analyte and the matrix, in which it was measured, can have a significant effect on the response in the MS; this is often referred to as ion suppression/enhancement effects. To compensate for these variations, correction of the response using internal standards is needed. These internal standards should have the same chemical behavior as the analyte but should be detected separately from the analyte of interest. The best internal standard for a certain metabolite is the stable isotopically-labeled (D, ^{13}C or ^{15}N) metabolite itself[78]. Once added to the sample the response of the (isotopically labeled) internal standard can be used for correction of different kinds of chemical and instrumental variations like sample treatment differences, pipetting errors, storage effects, ion suppression etc.. The ratio between the peak intensities of the analyte and internal standard gives an indication of the relative (to the selected internal standard) concentration of the analyte. Absolute quantification of the actual concentration levels (e.g. $\frac{\mu mol}{\mu l}$, $\frac{g}{kg}$ etc.) in all samples of the study can only be calculated if a calibration line for the metabolite of interest was included during measurement. For increased separation the MS is often hyphenated to a separation technique, e.g. gas chromatography (GC),

liquid chromatography (LC) or capillary electrophoresis (CE). In addition to an improved separation of analytes of interest, possible matrix effects and consequently ion suppression effects may be significantly reduced this way.

1.2 QUANTIFICATION

To get reliable quantitation (preferably absolute) the observed differences between the different analyzed samples should not be hampered by analytical variation and should be attributed only to real biological differences of interest. Consequently any further (data) analysis then solely can focus on interpreting these differences. The quantifiable response for a metabolite is the product of its concentration and a metabolite specific response factor. The response factor however, is affected by matrix effects which necessarily need to be minimized. The common ways to characterize these matrix effects are either by post column infusion methods or post-extraction spiking methods[23, 87]. Because the first method only characterizes the matrix effects qualitatively, the quantitative assessment using the second method is more common. With both methods however, the characterisation is biased towards a set of known metabolites only. In metabolomics where typically hundreds to thousands of (also unidentified) metabolites are measured, it is very uncommon to measure (internal) standards for each of these metabolites. This would be very laborious and thus expensive. In addition, it is not known a priori, which metabolites are of interest for the study at hand. As a consequence often platforms are used that cover a wide range of metabolites whose identity is not known in advance (so called untargeted platforms). In these cases usually at least one internal standard per class of metabolites is included to enable relative quantitation (e.g. on lipid per lipid class in lipidomics[53]).

The choice of a proper internal standard influences the estimated (relative) concentration of the compounds in question. Figure 4a shows the peak areas of L-Leucine and two internal standards that were added in replicated (GC-MS) measurements[8, 48] of over 100 identical reference samples (technical replicates [32], i.e. the complete analytical process rather than repeat injections of the same sample). Because the measurements concern the same sample, the ratio between the analyte and the internal standard (IS) should remain constant. The ratios of L-Leucine with the internal standards are plotted in Figure 4b. It is clear that correction with Leucine-D3 generates an almost constant value. However, it is also obvious that correction with a less suitable internal standard (in this case Phenylalanine-D5) can have a dramatic effect on the estimated relative concentration of L-Leucine.

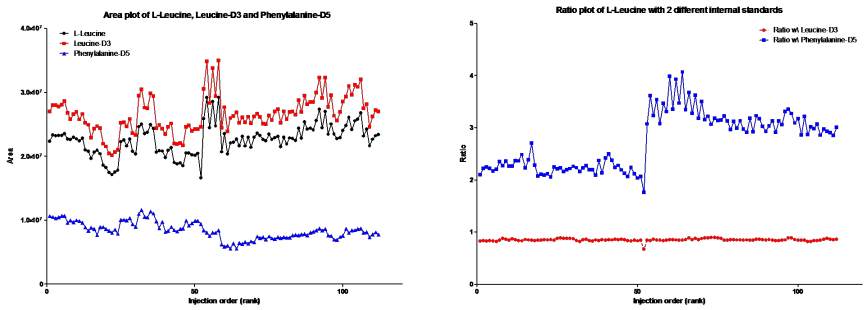

Figure 4: (a) The peak areas of L-Leucine and 2 internal standards for a series repeated measurements (112 technical replicates) of a QC sample. (b) The ratio plot of L-Leucine with each of the two internal standards.

If the ideal internal standard is not available or used, there are four levels of correction to consider to estimate the relative concentration: between analytes within one sample, between analytes over samples measured within one batch and, when many samples need to be measured that cannot be processed within one batch, between analytical batches of samples, and finally, when there is also a substantial time difference between measurements of sample sets, between studies correction. The common factor in all of these four levels is (acquisition) time and in specific all kinds of instrumental and environmental variations like matrix differences, sample degradation, different apparatus but also preprocessing/integration variation that have changed in this time. The challenge in metabolomics is to minimize these variations for as many as possible different metabolites. It is at this stage that metabolomics greatly benefits from statistics (e.g. experimental design [67], data analysis) but of course also from improved analytical sample preparation and analysis methods. With regards to analytical methods, one could think of using a different analytical setup (e.g. post-column infusion techniques [23]) that would quantify suppression effects for a whole range of metabolites but also other optimizations of experimental conditions like concentration levels of the added internal standard [104, 105, 10] can be considered. Statistically, a (mathematical) solution could be to construct virtual internal standards based on a (multivariate/linear) combination of internal standards to normalize the responses of unknown compounds. Finally, to improve comparison over analytical batches of samples and between studies appropriate reference samples could be used[54]. The choice which samples to use as a reference would be a clear result of the combined efforts in analytics and statistics.

1.3 INTEGRATION

Even if all analytical and instrumental settings are optimized, one issue in analyzing MS data that is often left untouched is the integration step itself. The principle to translate the area under the (unimodal and non-overlapping) curves to areas belonging to 2 different components as shown in Figure 5a is evident.

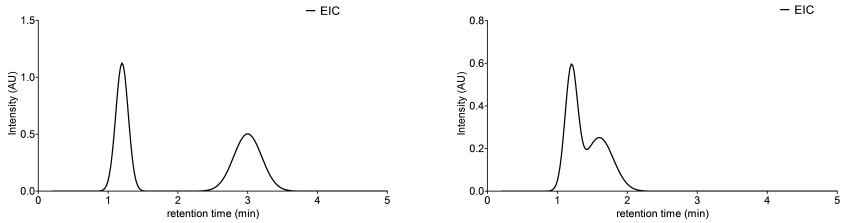

Figure 5: (a) Unimodal extracted ion chromatogram (EIC). (b) Bimodal EIC, (where) should the peaks be separated?

In case of bimodality or multi-modality curves (e.g. due to not fully separated isomers) things get more complicated and arbitrary decisions have to be made (Figure 5b); solutions are to calculate the sum of the total peak are under the curve, split them in the middle or try to fit the signal by (two or more) separate peaks (i.e. by deconvoluting them). Once a choice has been made the software has to be parameterized accordingly. Different (MS) vendors provide own software packages for integration and it is at this point where the different software packages show different outcomes for almost identical cases as depicted in Figure 6 (a and b).

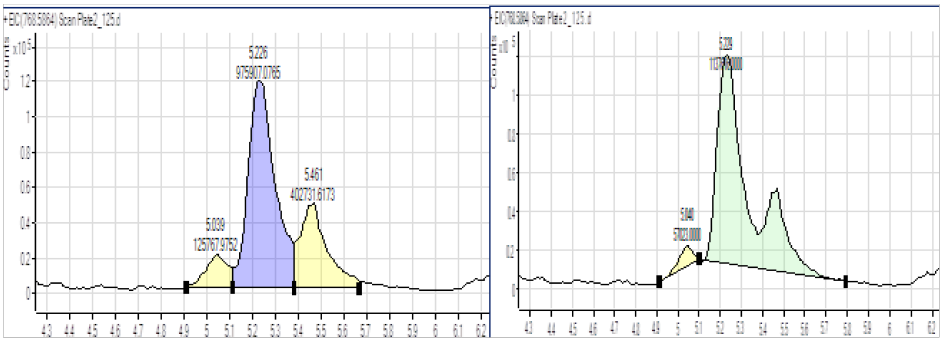

Figure 6: The unexpected behavior with automated integration software. (a) The peak is split in three separate peaks. (b) The two right peaks are combined.

In Figure 6 the same cases of multi-modal peaks are split in different ways using slightly different integration settings. Arguably in cases like this one option is to improve the chromatography but that is unfortunately not always possible, and is anyway time consuming. However, are such overlapping peaks really a problem? This depends on the type of research that is performed. If the aim is to extract known compounds/peaks only, the integration results can be validated by eye and manually adjusted if necessary, however, this is a time-intensive, and therefore expensive, process. If the approach is an untargeted profiling of analytes then there is no bias towards any specific analytes and consequently no (analyte) specific processing steps are there to configure. Visual optimization of integration parameters therefore is very difficult and

manual curation procedures as in targeted data processing is hardly feasible.

Despite the limited number of compounds reported and expensive manual data curation, targeted approaches are widely used. Obvious reasons are that the targeted metabolites/compounds are known which is very important for the data interpretation, and the possibility to quantify them (using internal standards and reference compounds) often with better precision and accuracy then in untargeted modes. To a large extent this is also due to the lack of appropriate software that would enable untargeted extraction and integration without introducing artifacts and errors. As a result, integration is often limited to a set of known metabolites (targets) only and in most cases vendor software is used for such targeted data processing.

1.4 STATISTICS AND DATA ANALYSIS

In metabolomics statistics are applied throughout the whole process of analyzing samples, from method development to data analysis. Experimental designs are applied to setup a study in such a way to minimize the number of experiments while retaining the maximum amount of information[67]. Repeated measurements of samples are used to statistically indicate whether or not the analytical platform functions within specification[37]. To this extent, often, for each specific metabolomics study, a pooled sample (a so-called Quality Control sample) from all samples of that study is created and repeatedly measured. As mentioned earlier, correction steps are necessary to compare metabolites between and over samples. Depending on the type of sample, the way it was measured etc., a whole range of statistically data pretreatment (normalization) methods are offered to improve ultimately the biological interpretation of the data[127]. Actually most, if not all, statistics (in metabolomics) are performed to remove/indicate analytical variation in metabolites (features) that are measured. Those features that do not meet the pre-defined criteria are usually removed from the dataset and further data analysis/interpretation is continued with a smaller set of reliable metabolites. This removal does not necessarily improve biological interpretation but the complexity of follow-up (data) analysis can be reduced considerably.

In univariate statistics the focus is on one variable at a time and the results are relatively easy to interpret from a statistical point of view (e.g. the effect of the variable is significant or not using T-tests[89]). As a consequence univariate statistics methods -are widely accepted, especially in clinical settings[147, 100]. Because changes in biological samples are often multifactorial[147, 100], metabolomics data should be analyzed using multivariate statistics as well. In contrast to univariate statistics multivariate statistics focusses on simultaneously analyzing a set of variables. The (relative) importance of the individual variables in answering the biological question is not always that straightforward and easy to determine but the multivariate profiles however, often do reveal important variables that would have not appeared relevant based on univariate statistics only. After proper quantification, principal component analysis (PCA) is often used to do pattern recognition and visualize observed group differences[86] and methods like partial least squares − discriminant

analysis (PLS-DA)[142] are commonly used to relate these differences to specific metabolites. Using statistical modeling of properly quantified metabolites, (multivariate) metabolic networks can even be inferred[44]. Because of the limited number of samples in comparison to the huge amount of variables (e.g. metabolites) that are measured, multivariate models easily lead to overfitted results (i.e. perfect fits are found, but the predictive power of the model is limited). The results are hugely aided by variable selection methods to select the important from the less important variables and cross-validation and permutation[142] procedures to prevent this overfitting when building predictive multivariate models.

1.5 SCOPE AND OUTLINE OF THIS THESIS

In the previous paragraphs some typical challenges were discussed that researchers are faced with when handling data from metabolomics studies using untargeted mass spectrometry based data. The aim of this thesis was to develop concepts and methods to extract qualitative and quantitative information about metabolites from untargeted mass spectrometric data. For this, different methods were developed to obtain quantitative metabolite data in large studies using GC-MS, LC-high resolution MS (HR-MS) and direct infusion high resolution mass spectrometry. The different methods address different parts in the metabolomics workflow, i.e. data -acquisition, data pre-processing up to data-analysis.

As the performance of analytical systems can vary, different methods of normalization to improve quantification for known and unknown compounds were developed. In **Chapter 2** it is demonstrated that for (relative) quantification of metabolites in GC-MS metabolomics studies, in the absence of matched stable isotopes, per metabolite normalization based on a single internal standard is not enough to correct for analytical batch-to-batch differences. This is especially troublesome in large scale metabolomics studies where many samples need to be measured and consequently many analytical batches are needed. Furthermore, even within a single analytical batch a clear trend in the response for specific metabolites was observed. A statistical procedure based on repetitive measurements of identical samples (i.e. technical replicates) is suggested that corrects for these batch-to-batch differences even for metabolites without a proper internal standard.

In the search for biomarkers for Diabetic Kidney Disease (DKD) in **Chapter 3** LC-MS data of urine samples of an epidemiological study were analyzed. Data acquisition was for that data set unfortunately suboptimal, and various variations in the data were present making (relative) quantification of this untargeted data set difficult. Still, after extensive data preprocessing, a clean data set was obtained suitable for data analysis. It was shown that multivariate statistical modeling was advantageous over univariate modeling for the discovery of biomarkers for this data set. Penalized logistic regression models were used to create a predictive model. Double-cross validation was used to reveal potential new biomarkers.

In **Chapter 4** a method has been developed and demonstrated for the processing of another type of very complex metabolomics data, i.e. metabolomics data obtained by direct infusion mass spectrometry. It was demonstrated that with the preprocessing method that was developed, biological relevant results, i.e. the characterization of different development stages of zebrafish embryos, could be extracted from these very complex metabolomics data. Feature identification was solely based on accurate mass and therefore the samples were recorded with a very high mass resolution. The method developed was based on the binning tools developed for LC-MS (Chapter 5) by aligning the masses over samples which enabled further automated data analysis. Internal standard correction for the unknown features was based on the same strategy as described in Chapter 2. In the absence of quality control samples however, the relative standard deviation (RSD) was calculated using replicated measurements.

The integration problems that were observed during pre-processing of untargeted LC-MS data from earlier experiments (including those reported in Chapter 3), led to the awareness of the lack of good software to integrate peaks in such data sets. The freely available software options required much expertise to configure and were not robust enough to quantify metabolites present at low intensities with good precision and accuracy. In **Chapter 5** therefore a new approach was introduced to integrate samples acquired using LC-time-of-flight-MS. The samples were automatically processed one-by-one to facilitate (future) parallel processing. With only a few parameters that need to be set the user interaction is kept to a minimum, but at the same time obtaining reliable quantitative data on peak areas of known and unknown metabolites.