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).
The handle [http://hdl.handle.net/1887/25808](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
SUMMARY AND FUTURE PERSPECTIVES

The aim of this thesis was to develop concepts and methods to extract qualitative and quantitative information about metabolites from untargeted mass spectrometric data of biological samples. Several typical challenges in data handling were addressed that prevent a straightforward interpretation (data analysis) of the data acquired with different types of mass spectrometric-based metabolomics methods (GC-MS, LC-MS, CE-MS or DI-MS) methods. The critical parameters causing variation in quantitative results were identified and studied at different stages in the metabolomics workflow such as data acquisition (Chapter 2), data pre-processing (Chapters 2, 4 and 5) and data analysis (Chapters 3 and 5). Different methods and concepts were developed to address these and to improve the quantitation of metabolites and the comparison between metabolite data in different samples of the same study measured at different moments or between studies. The methods developed focused on improved normalization, data pre-processing of untargeted analysis and data pre-processing of high resolution direct infusion mass spectrometry data. Furthermore it was demonstrated that even for metabolomic studies with few samples cross-validation of multivariate models can be very time consuming and parallel implementation on a (large) cluster of computers is the way to make such computations feasible.

All methods were developed in such a way that they can be used as an automated module within the data processing pipeline. In Chapter 2 it was shown that quantification of metabolites in metabolomics studies greatly can be improved using single point calibration. The abundance of each metabolite is related to a predefined amount of internal standard added to each sample. This one-point calibration is indispensable when large-scale metabolomics studies are performed, in which both within and between batch differences become a problem due to instrumental and environmental changes during the measurements of the large number of samples. By repeated measurements of a representative sample for the study, usually pooled Quality Control (QC) samples, we demonstrated that the relative standard deviation (RSD) of individual metabolites in these QC samples before and after internal standard correction is a good measure to find the best match between a given metabolite and a set of internal standards that were spiked in the sample. This is a practical alternative to using a separate (isotope labeled) internal standard for each metabolite, which is often not feasible due to high costs and/or limited availability of such standards. It was shown that two types of QC samples are required whereby the first type, the calibration QC samples, is used to perform a one-point calibration, and the second type, validation QC samples, is used
to assess how well the calibration procedure improves the data quality. As a result, the biological variation in the study samples becomes more apparent, and more meaningful statistical models can be built.

In Chapter 3 it was demonstrated that based on untargeted LC-MS measurements of urine samples a statistically significant multivariate model could be constructed to distinguish between progressive and non-progressive subjects within the normal urinary albumin excretion rate (AER) group with 75% accuracy. The metabolic profile defined by the multivariate model included all of the measured compounds that showed a univariate significant difference between the two groups. The profile, however, also included metabolites that did not show a univariate significant difference and emphasizes the additional benefit of multivariate statistics over univariate statistics alone in preventing overlooking candidate biomarkers. Key for multivariate modeling however is proper model validation and permutation testing.

We demonstrated the use of rapid metabolic fingerprinting for rapid determination of metabolic changes in Chapter 4 and we were able to distinguish metabolic profiles of early developmental stages of zebrafish embryos using High Resolution Direct Infusion Mass Spectrometry (HR-DI-MS). Actually, in this project data preprocessing was not the focus of the project at the beginning, but the lack of automated data pre-processing of this type of data initiated the development of the method described in this chapter. The huge number of features that were generated in a single mass spectrum made clear that some kind of unique (virtual) reference point (e.g. a feature appearing in all samples) over the samples was needed. After careful data preprocessing and analysis, we were able to isolate 102 features that showed consistent behavior within each developmental stage. Principal Component Analysis revealed that early development stages of zebrafish embryos could be differentiated from each other. In total 27 out of the 102 features were (putatively) identified. Several observed trends of these putatively identified metabolites were supported by previous publications. But more importantly, with our method several new features were discovered as being relevant during early development of zebrafish embryos. It could be concluded that HR-DI-MS is suitable for rapid metabolic profiling on zebrafish embryos. However, to improve robustness and obtain more high-quality features better, fast but appropriate sample preparation methods are required.

In Chapter 5 a new method to integrate high resolution full scan profiling LC-MS data in an untargeted manner is introduced. To demonstrate the effectiveness of the strategy of only comparing feature-sets over samples we used complex lipidomics full scan profiling LC-MS data. The automatically integrated areas for a set of known target lipids were compared to those obtained by optimized and manually controlled quantification using vendor software, i.e. which were considered as the reference data. Very high correlations with the reference data were observed which was impressive since the integration parameters for the different lipids (e.g. combining/splitting of isomers) in the vendor software have been highly optimized over a period of years. The untargeted method extracted at least 10 times more feature-sets than the known
target lipids and PLS-DA models based on either the new found feature-sets or on the extracted known targets performed equally well. Selectivity ratios however, showed that the most important discriminating feature-sets were contained in the set of higher abundant unknown feature-sets, confirming the potential biological relevance of these feature-sets found and thus indicating the added value of untargeted integration, as these allow also the detection of metabolites and lipids at low concentrations, and include also so far unknown metabolites. The proposed integration method requires only a very limited amount of expert knowledge and is fully automated with almost no user interaction which makes it a perfect candidate for inclusion in a data pre-processing pipeline. Because the extraction and integration method is implemented on a per-sample base, the method is highly scalable when more computers/processors are available. We envision this method to be extended to also facilitate GC-MS feature extraction but also enable automated CE-MS data analysis that suffers from huge migration time shifts between samples.

With respect to the title of this thesis, one may ask the question whether quantification of untargeted mass spectrometry data was significantly improved by the methods described in this thesis, and are these methods suitable for metabolomics research project? The answer is complex as the developed methods indeed improved the quantification of metabolites and allowed to identify and deal with the sources of analytical variation, but more developments are required.

The inclusion of multiple internal standards in untargeted profiling methods is essential and it should always be investigated which internal standards to use. Often there is a high correlation between the added internal standards. This is partly due to the way the samples are prepared (i.e. adding a mix of internal standards as one solution). Ion suppression can jeopardize quantification. The experimental setup to qualitatively screen for regions that suffer from matrix effects using post column infusion is very interesting. Infusing a mixture of reference signals could then act as calibration points and correct for the amount of suppression at that point in time, i.e. acting as internal standards.

The huge number of additional features that was generated using the untargeted integration method also introduces a new challenge: what is the right strategy to analyze these features, which feature-set is good, which is not? To test this, often, like we did, the reproducibility in repeated measurement of pooled quality control (QC) samples is used. If however the focus is on biomarker discovery such a strategy is riskful. The nature of a pool of QC samples is that it is an average of all metabolites that can be expected. But what if the (unknown) biomarker that we are looking for is too much diluted in this average sample? If the concentration of a biomarker is close to the quantification limit, and if it is only present in one of several classes of a (clinical) study, the concentration in the pooled QC sample may be so low that the RSD values of repeated measurements in the QC samples are unacceptable high and the biomarker will be excluded for further data analysis. In cases like these, a pooled reference sample per sample class group would be beneficial.
RSDs can then be determined per sample class and data-analysis would be enriched with metabolites that otherwise would not have passed the criteria. Consequently the standard data acquisition protocol should be adjusted accordingly for future metabolomic studies by including multiple class-specific QC samples. Experiments should quantify/qualify this dilution effect and assess whether pooled QC samples per control group are more representative than overall pooled QC samples for all metabolites in both groups. In these experiments the response of the QC pool per group(s) should be compared to the response of the combined (regular) pooled QC sample over a large number of measurements to mimic the experimental variations of a large metabolomics study as much as possible.

The importance of the QC samples becomes more and more apparent but the limited amount of material often prohibits re-measurement in different studies but sometimes there is not even enough material available to create an adequate QC pool for the original study. In such cases it would be of much interest to evaluate the response of different types of reference samples. Do these reference samples behave adequately for QC monitoring and QC correction and/or would they act as good validation samples to test the QC correction and IS correction steps? It would be even more interesting to evaluate if these references samples could be used as transfer samples to create so-called transfer models and enable direct comparisons between different studies.

In Chapter 5 feature-sets like molecular adducts that were related to known lipids were explicitly left out not to influence the model prediction. But what if they were also included? It is known that in some matrices (e.g. different samples, experimental conditions) adducts are more easily created then in others. It could be expected that the sum of all compound related masses could improve the quantitative comparison between the different samples and minimize matrix effects, another hypothesis worthwhile to test in the future.

In conclusion, significant progress in the data processing of MS-based metabolomics data was achieved in this thesis, but much progress still has to be and will be made in this field. The ever increasing computing power enables ever more complex data extraction, data processing and data analysis procedures. The popularity of untargeted methods is likely to increase as a consequence of the development of new software, complex deconvolution algorithms, higher mass accuracies and reference databases. And such untargeted methods allow the detection of a target list of metabolites. But even though fewer compounds are measured and/or reported, targeted methods will always have its advantages (e.g. less processing time), especially when high throughput of samples is of concern. The success of an analytical platform hugely depends on the ability of the subsequent data (pre)processing steps to obtain high quality data. This dependency will always be there but new data (pre)processing techniques may also inspire technical changes and/or different analytical setups. In conclusion, metabolomics research should be approached as an integrative effort combining knowledge on the biological question, sample processing and data acquisition and data processing, and subsequent data analysis, to ultimately answer biological questions.