Dissertation

IPO: A Tool for automated Optimization of XCMS Parameters

submitted by

Dipl.-Ing. Gunnar LIBISELLER, Bakk.rer.soc.oec.

for the Academic Degree of

Doctor of Medical Science (Dr. scient. med.)

at the

Medical University of Graz

Department of Internal Medicine

Division of Endocrinology and Metabolism

under the Supervision of

Dipl.-Ing. Dr. Frank Michael Sinner

2015

Declaration

I hereby declare that this dissertation is my own original work and that I have fully acknowledged by name all of those individuals and organisations that have contributed to the research for this dissertation. Due acknowledgement has been made in the text to all other material used. Throughout this dissertation and in all related publications I followed the guidelines of ‘Good Scientific Practice’.

Graz, July 2015

Please note that parts of this dissertation are already published:

Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics [Internet]. 2015;16:118. Available from: http://www.biomedcentral.com/1471-2105/16/118

Acknowledgements

This thesis is the result of my work at the research group Bioanalysis and Metabolomics of the HEALTH institute of Joanneum Research Forschungsgesellschaft mbH.

First I want to thank Mag. Dr. Christoph Magnes for giving me the opportunity to write this thesis and for his guidance and mentorship. Furthermore I'm very grateful to the members of my committee Dipl.-Ing. Dr. Frank Sinner, Prof. Dr. Thomas Pieber and Dr. Harald Sourij. Additionally my thanks go to the co-authors of the IPO publication Mag[a]. Michaela Dvorzak, Dipl.-Ing. Ulrike Kleb, Edgar Gander, Dr. Tobias Eisenberg, Prof. Dr. Frank Madeo, Dr. Steffen Neumann and Gert Trausinger. Michaela and Ulrike started the work for parameter optimization which I was allowed to continue. Edgar, Tobias, Frank and Gert provided valuable data for testing the approach and Steffen greatly helped to improve the usability of the developed tool. I'm also very grateful to Dipl.-Ing. Dr. Beate Boulgaropoulos and Mag[a]. Dr. Selma Mautner for the countless times they read through and revised the publication. Additionally I thank my PhD-college MMag[a]. Sophie Narath for having a sympathetic ear regarding all kind of arising PhD-issues.

Last but not least I want to thank my wife, my parents and my brother for just being who they are and of course all my friends.

Table of Contents

Abbreviations and Definitions

Abbreviation	Description
ARTS	Average retention time score
BBD	Box-Behnken design
CCD	Central-Composite design
centWave	An algorithm for peak detection provided by XCMS
CWT	Continuous wavelet transform
density	An algorithm for grouping provided by XCMS
DoE	Design of Experiments
EIBPC	Extracted ion base-peak chromatogram
ESI	Electro spray ionization
GRTS	Group retention time score
GS	Grouping score
HILIC	Hydrophilic interaction chromatography
HPLC	High performance liquid chromatography
HRMS	High resolution mass spectrometry
KDE	Kernel density estimator
LC	Liquid chromatography
LIP	Low intensity peaks
loess	Locally weighted scatterplot smoothing and also an algorithm for retention time correction provided by XCMS
m/z	Mass-to-charge ratio
matchedFilter	An algorithm for peak detection provided by XCMS
MS	Mass spectrometry
mzXML	open data format for LCMS data
nESI	Negative ESI

NMR	Nuclear magnetic resonance
obiwarp	An algorithm for retention time correction provided by XCMS
pESI	Positive ESI
Plakett-Burman design	A fractional factorial design
PPS	Peak picking score
QC	Quality control
RCS	Retention time correction score
RGTV	Retention time correction and grouping target value
ROI	Region of interest
RP	Reliable peaks
rsm	R-package for response surface model
TIC	Total ion count
UHPLC	Ultra high performance liquid chromatography
XCMS	R package for profiling LC-MS data

List of Figures

List of Tables

Zusammenfassung

Untargeted Metabolomics hat das Ziel möglichst alle molekularen Substanzen in einer biologischen Probe zu erfassen und somit einen Fingerabdruck der Probe zu generieren. Dadurch erhofft man sich Rückschlüsse auf Stoffwechselvorgänge ziehen zu können und in weiterer Folge die Einflüsse von Stoffwechselprodukten auf Krankheiten und umgekehrt auch die Auswirkungen von Krankheiten auf Stoffwechselprodukte feststellen zu können. Zur Erfassung dieser Stoffwechselprodukte oder Metaboliten kommen moderne, hochauflösende Massenspektrometer zum Einsatz. Massenspektrometrie gekoppelt mit Gas- oder Flüssigkeitschromatografie erzeugt eine große Menge dreidimensionaler Daten. Durch die enorme Menge an Daten ist eine manuelle Auswertung nicht durchführbar. Daher wurden Programme entwickelt, die automatisch Rauschen filtern, Signale in den Daten identifizieren, korrigieren und diese Signale zu aussagekräftigen metabolischen Features zusammenführen. Um auf die unterschiedlichsten Arten von Daten möglichst flexibel einstellbar sein zu können, verfügen diese Softwareprodukte über verschiedene Parameter. Dabei ist die richtige Wahl der Parametereinstellungen nicht trivial und hat großen Einfluss auf die Verlässlichkeit der produzierten Daten. Bisher wurden nur wenige und auch nur theoretische Ansätze zur automatischen Optimierung dieser Parametereinstellungen publiziert. Bisher wurde noch kein Programm entwickelt welches diese Optimierung automatisch vornimmt und die bestmöglichen Einstellungen liefert.

Diese Dissertation hatte daher zum Ziel eine Software zu implementieren welche automatisch eine Optimierung der Parametereinstellungen für beliebige Daten durchführt. Dabei sollten die Parameter der weit verbreiteten Open Source Software XCMS optimiert werden. In einem ersten Schritt wurde zur Optimierung der Parameter der Detektion von Signalen eine neue Zielgröße entwickelt. Diese basierte auf der Identifikation von Signalen die von natürlichen, stabilen Kohlenstoffisotopen stammen. Diese Signale, sowie die entsprechenden Signale ohne Kohlenstoffisotop, wurden als verlässlich definiert und führten zu Berechnung der Zielgröße. In einem nächsten Schritt wurde eine weitere Zielgröße definiert die sowohl die mittleren,

relativen Retentionszeitverschiebungen aller metabolischen Features reduziert und gleichzeitig die Anzahl verlässlicher metabolischer Features steigert. Die verschiedenen Parametereinstellungen des Parameteroptimierungsverfahrens werden mit statistischer Versuchsplanung ausgewählt und mittels Response Surface Modellen evaluiert. Durch die effiziente Versuchsplanung und die Aufteilung des Optimierungsproblems in zwei Schritte wurde die Anzahl der notwendigen Experimente auf ein Minimum reduziert was die Dauer der Optimierung stark verkürzte.

Die entwickelte Software wurde an vier verschiedenen Datensätzen getestet welche aus unterschiedlichen Probenarten stammen und mit unterschiedlichen Geräten und Methoden gemessen wurden. Bei allen Datensätzen konnte eine signifikante Steigerung der Zielgrößen und somit der Verlässlichkeit der Daten beobachtet werden.

Die entwickelte Software wurde im Journal BMC Bioinformatics publiziert und ist unter www.github.com/glibiseller/IPO frei erhältlich.

Abstract

Untargeted metabolomics aims to detect as many low-molecular-weight substances in a biological sample as possible and thereby generates a fingerprint of the sample. The fingerprint enables conclusions about the metabolic processes to be drawn and subsequently to determine the impact of metabolites on specific diseases and also the influence diseases have on metabolites. To detect these metabolites modern, high-resolution mass spectrometers are used. When they are coupled to gas or liquid chromatography a large amount of data is generated. Due to the enormous amount of data manual processing is not feasible. Therefore, programs have been developed to automatically filter noise, identify and correct signals in the data and merge the signals into meaningful metabolic features. In order to be as flexibly as possible and adaptable to various types of data these software products provide different parameters. The right choice of parameter settings is not trivial and has great influence on the reliability of the resulting data. So far, only a few theoretical approaches for automated optimization of these parameter settings have been published. A program that automatically performs parameter optimization and provides the best possible settings has not yet been developed.

Therefore the aim of this thesis was to implement a software tool that automatically performs an optimization of parameter settings for any type of data. The parameters of the prevailing open-source software XCMS were to be optimized. To gain best possible parameters of peak detection in a first step a new score was developed. This score was based on the identification of signals derived from natural, stable carbon isotopes. These signals and the corresponding signals without carbon isotope were defined as reliable and led to the calculation of a score. In a next step, an additional score was defined which aimed to decrease the mean, relative retention time shifts within metabolic features while also increasing the number of reliable metabolic features. The various parameter settings were chosen and evaluated using a design of experiment approach. By using efficient design of experiments and splitting the optimization approach into two separate steps the number of necessary experiments

was reduced to a minimum which consequently reduces the time needed for the optimization.

The developed software has been tested on four different datasets originating from different types of samples which were measured with different devices and methods. All datasets showed a significant increase in the scores and thus an increased reliability of the data was observed.

The developed software tool was published in the journal BMC Bioinformatics and is freely available from www.github.com/glibiseller/IPO.

1. Introduction

Untargeted metabolomics screens biological samples with the aim to find new compounds and to identify potential biomarker molecules. The most used technology in metabolomics is mass spectrometry (1). When using state of the art high resolution mass spectrometry coupled to liquid chromatography (LC-HRMS) devices for biomarker research, such methods produce a huge amount of data. This amount cannot be interpreted as it is and therefore automatic data processing is indispensable. The metabolomics workflow consists of several steps (**Figure 1**). Biological samples are acquired and prepared to be measureable by devices like MS or nuclear magnetic resonance (NMR). The achieved raw data is processed and the result statistically analysed. Potential biomarkers are identified with targeted MSMS analysis and annotated. In the last step of the workflow the resulting data is biologically interpreted. This thesis focuses on the data pre-processing steps peak picking, retention time correction and grouping. In the peak picking step veritable signals from substances are identified and noise is filtered. Run to run deviations in the chromatographic axis are corrected in the retention time correction step. Grouping is the process of finding and combining peaks measured in different samples but with similar retention times and mass-to-charge ratios to produce peak groups. Due to the automated processing, creating reliable results is very important. Recently, the tool MetExtract has been presented (2,3). It compares samples labeled with stable carbon isotopes and unlabeled samples to find reliable peaks. This tool increases the selectivity of compounds with biological origin, performs peak group reduction to reliable groups only and can even assess molecular structures of measured substances. However, the tool is only applicable to samples that can be labeled. Also the labeling step is very time and cost intensive and is a major disadvantage of this method.

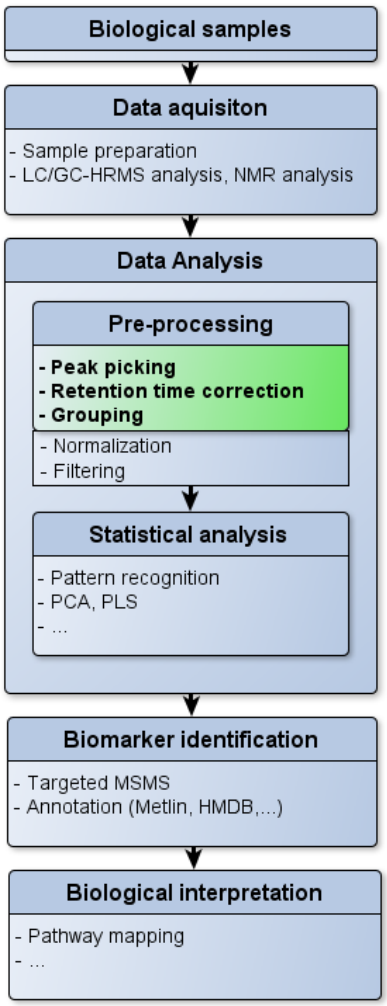


Figure 1: The untargeted metabolomics workflow is illustrated. Biological samples have to be prepared and measured. The achieved raw data is analyzed and the results used to identify potential biomarkers which are then interpreted. This thesis focuses on the steps highlighted in green.

An additional number of tools exist which do not rely on labeling (4–18). In order to adapt these tools to different LC methods, different types of mass spectrometers as well as different types of samples various parameters are provided. These parameters greatly affect the reliability of the resulting data. With the aim to achieve optimized parameters various approaches have been developed (19,20). One parameter optimization approach adds additional information to the data by creating a dilution series of a pooled sample (19). Peak groups highly correlating with the

dilution series are defined as reliable whereas peak groups which do not correlate are defined as unreliable. A reliability index is calculated by the ratio of the squared number of reliable peaks to unreliable peaks. The tested parameter settings are achieved by a design of experiments (DoE) approach based on Central-Composite designs. This parameter optimization approach provides on the one hand quality evaluation of the resulting optimization but on the other hand it is time consuming because all parameters are unbiasedly optimized.

To target the time consuming nature of this approach, Zheng, H et al. (20) accelerated it by first applying a screening step prior to the actual optimization. Screening experiments are usually performed in the first stage of an optimization process with the purpose of identifying the parameters that have large effects on a target variable. The target variable in their case is the calculated reliability index. Their screening step is based on a Plakett-Burman design. This design is a fractional factorial designs which sets only two levels for each parameter and therefore allows the user to accomplish a screening step with relatively few experiments. Two levels denote that two different settings are tested for each parameter. After the screening step only parameters with a significant positive influence on the target value are optimized following Eliasson et al.'s approach and leads to a significantly decreased overall optimization time. The reliability of the resulting data was further increased by using different filters. However, potential important parameters might not be optimized because they have a significant negative influence or they are not significant enough due to the parameter levels used for testing in the screening step.

Other approaches aim to increase the reliability of the metabolic data itself (21,22). xMSanalyzer, a tool to increase the reliability of untargeted metabolomic data was published by Uppal et al. (21). The tool is a processing pipeline which supports apLCMS and XCMS. In a first step, peaks are found by using different parameter settings which results in multiple datasets. If replicate samples are available in the dataset, peaks are only forwarded to the next step if they are found in all replicates. Peaks from all datasets are then merged together and unique peaks as well as overlapping peaks are identified. This approach increases the detection of low abundance metabolites and by using replicates the reliability of detected peaks is

increased. However, the tested settings have to be chosen by the user. Inexperienced XCMS users may not be able to apply reasonable settings. Also the best possible settings may be far from the tested ones. Therefore, it can be concluded that xMSanalyzer increases the reliability of untargeted metabolomic data but is not a parameter optimization approach in the proper sense. The same is true for the quality control approach published by Brodsky et al. which uses replicate samples for evaluation of the peak picking quality and for applying data normalization (22).

2. Material and Methods

This chapter gives a basic overview of the characteristics of the data generated by a LC-HRMS device. A very detailed description of the XCMS functions and their parameters for processing untargeted metabolomic data is provided. Additionally the newly developed optimization approach is described and the datasets used for the evaluation of the optimization approach are presented.

2.1. Chromatography and mass spectrometry

To understand the challenge of untargeted metabolomics data processing it is necessary to understand the characteristics of the processed data first. State of the art high performance LC-HRMS (HPLC-HRMS) devices can measure thousands of substances within a single biological sample. Their selectivity and sensitivity make them the perfect tool for metabolic profiling (23). A HRMS can detect multiple masses with each scan (**Figure 2**). However, the HRMS device does not measure a mass directly but a specific mass-to-charge ratio (m/z). This implies that an MS can only detect charged substances. The higher the concentration of a substance is, the higher the signal the MS measures. The data shown in **Figure 2** was recorded in profile mode. The picture at the top shows the complete mass range from 70 m/z to 600 m/z. The profile mode can be especially seen in the middle spectrum where the mass peak spans an m/z range from 144.05 to 144.075 creating a peak in the m/z dimension. The lower picture shows the same m/z segment but in a different presentation method. If the data were not profile but centroid data, only the green stick in the lower image would be recorded making the mass peak indefinitely small and therefore not creating a peak in the m/z dimension.

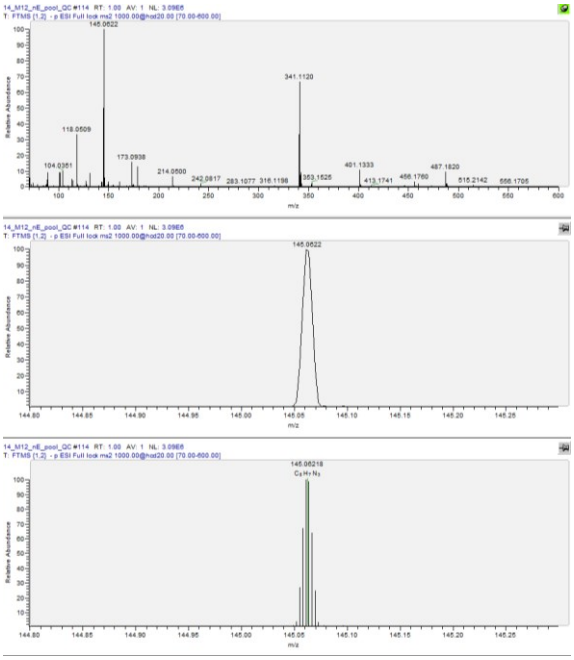


Figure 2: Mass spectrum: The upper image illustrates a complete mass spectrum from m/z 70 to m/z 600 of a single scan. The image in the middle shows a close up of the mass range 144.4 to 145.5. The lower spectrum shows the same but instead of a profile it shows the detected mass signals as sticks.

Preceding the mass spectrometer by chromatography introduces a third dimension, the retention time. The chromatography is used to separate substances with the same mass on the chromatographic axis. The mass spectrometer produces spectra on each time point of the chromatography resulting in three dimensional data (**Figure 3**). The three dimensions are m/z, retention time and intensity.

The image on top of **Figure 4** shows the whole LC-HRMS data from the chromatographic perspective. The middle image illustrates the whole chromatogram but is reduced to the m/z range from 145.0593 to 145.0651. The lower image shows the same m/z range but only retention time from 0 to 1.5 minutes. In the image at the bottom a chromatographic peak is well perceptible which is the kind of peak untargeted metabolomics aims to identify.

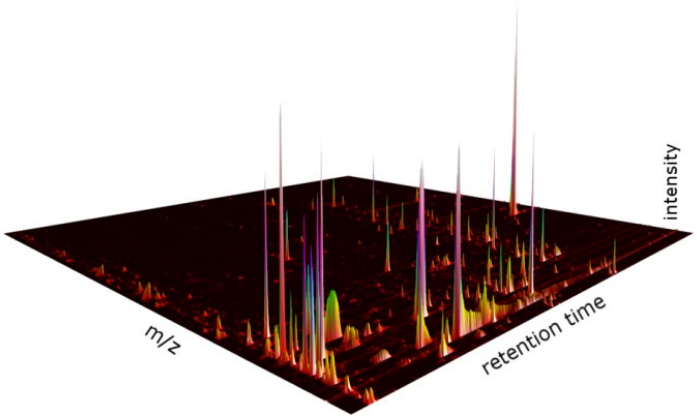


Figure 3: LC-HRMS data of one measurement. The data is three dimensional hence each measurement point is defined by m/z, retention time and intensity.

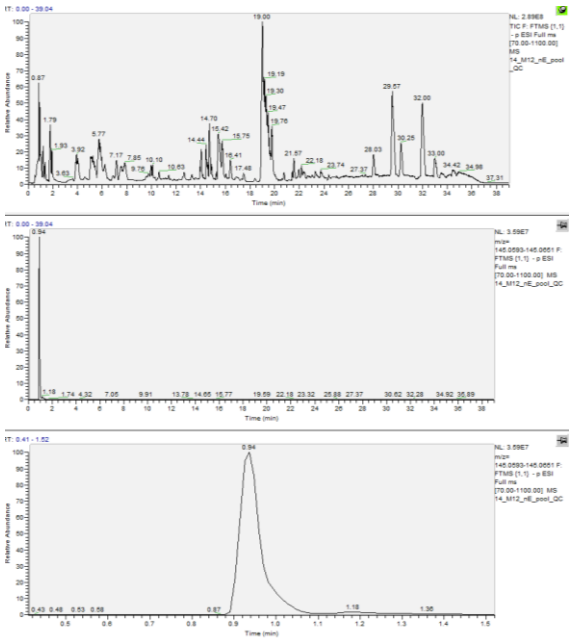


Figure 4: Different ranges of a chromatogram are shown. The upper chromatogram shows the whole m/z range displaying the highest total ion count (TIC) on each time point. The middle chromatogram only shows a very small m/z range. The chromatogram at the bottom show a specific m/z and retention time range.

2.2. XCMS

XCMS (4,5) is an open source R-package used for untargeted metabolomics based on LC-MS. The software can read and process different open data formats like mzXML (24), mzML (25–27), mzData (28) and netcdf. Untargeted metabolomic data processing consists of the steps peak detection, retention time correction, grouping and filling in data of missing peaks.

2.2.1. Peak detection methods

The purpose of peak detection is to find veritable chromatographic peaks representing biological substances and to distinguish them from noise. XCMS provides several methods for peak detection of which all have their pros and cons.

2.2.1.1. ‘matchedFilter’

The ‘matchedFilter’ algorithm can handle profile and centroided data (4,29). It can adapt to specific measurement methods by setting its parameters accordingly (Table 1). The algorithm starts by cutting the LC-MS data into m/z slices which is called binning. The wide of these slices can be defined by the parameter ‘step’. To big step size can lead to smaller peaks being swallowed by big peaks with a similar m/z value, whereas to small step size can lead to jagged peaks which are not recognized as peaks anymore. Multiple mass signals might be detected at each retention time point of these slices. By taking the maximum signal at each time point an extracted ion base-peak chromatogram (EIBPC) is generated. The algorithm then works on each of these EIBPCs individually. A second derivate Gaussian model is created and used for matched filtration on the EIBPC. The standard deviation of the Gaussian can be defined either by the parameter ‘sigma’ directly or indirectly as full wide at half maximum by the parameter ‘fwhm’. The detected peaks are then selected using a signal to noise cutoff which can be specified by the parameter ‘snthresh’. The amount of peaks that should be detected in each EIBPC is limited by the parameter ‘max’.

Issues may occur at the m/z boundaries of the EIBPCs. A peak may be wider as the m/z slice and transcend multiple slices being detected several times in adjacent EIBPCs. The parameter ‘steps’ defines how many adjacent slices should be looked into to find such peaks. Another issue could steam from very sharp mass peaks or as an extreme example centroided data. The signals could alternate between two EIBPCs. However, the EIBPCs can overlap at the boundaries to counter that. The parameter ‘mzdiff’ specifies how big this overlap should be. For data from HRMS the default values for ‘step’ and ‘mzdiff’ are too big and should be reduced accordingly. If not the whole mass- or retention time range of the LC-MS data is of interest, a smaller range in both axes can be specified using the parameter ‘scanrange’.

Table 1: Parameters and default values of the peak detection method ‘matchedFilter’

Parameter	Default value
fwhm	30
sigma	fwhm / 2.3548
max	5
snthresh	10
step	0.1
steps	3
mzdiff	0.8 – step * steps
scanrange	numeric()

2.2.1.2. ‘centWave’

This algorithm for peak detection is very well suited for centroided data from HRMS (30). It is not well suited for profile data however it takes into account the problems of the binning used by the ‘matchedFilter’ algorithm. The LC-HRMS data is not cut into m/z slices but regions of interest (ROIs) are detected. In a first step all m/z values

from the first scan are added to a ROI list. Then for all next scans m/z values are added to one of these ROIs where it has a deviation of less than specified by the parameter 'ppm' of the ROI's mean m/z. If the deviation is bigger for all existing ROIs a new ROI is added to the ROI list. Only centroids with an intensity larger than defined by the parameter 'noise' are used for ROI detection. The parameter 'prefilter' allows to early discarding ROIs of low intensity by setting two values, k and I. A ROI is only valid if it has at least k consecutive centroids with intensity I. In the remaining ROIs peaks are then detected by using Continuous Wavelet Transform (CWT) on a normalized second derivate Gaussian. The parameter 'peakwidth' defines a minimum and maximum wide of expected peaks. In combination with the average time between scans the value p_{min} is calculated. ROIs with less centroids than p_{min} are not processed. Only peaks with a signal to noise ratio greater than the setting defined by the parameter 'snthresh' are valid. The parameter 'mzdiff' defines the minimum m/z difference of peaks with the same retention time. These peaks may even overlap if 'mzdiff' has a negative value. A specific range for peak detection can be specified by the parameter 'scanrange'. The ROI detection can be skipped if ROIs are defined by the parameter 'ROI.list'.

Table 2: Parameters and default values of the 'centWave' peak detection method

Parameter	Default value
ppm	25
noise	0
prefilter	c(3,100)
peakwidth	c(20,50)
snthresh	10
mzdiff	-0.001
ROI.list	list()
scanrange	numeric()

2.2.2. Grouping methods

Grouping is the process of combining peaks from different samples but with similar mass-to-charge ratios and similar retention times to peak groups which are also called metabolic features. These metabolic features are more reliable than single peaks since the same peak was measured in different samples reducing the risk of noise being falsely interpreted as peak. Different methods for grouping exist.

2.2.2.1. ‘density’

The grouping method ‘density’ cuts the spectra into m/z slices (4). The width of these m/z slices is defined by the parameter ‘mzwid’. To prevent peak groups from being cut into two groups at the half width of ‘mzwid’ additional slices are made. Therefore peak groups may be detected twice which is later corrected in a post-processing step. In each of the m/z slices peak groups are searched individually. The retention times of all peaks are combined to a peak density chromatogram. This peak density chromatogram is smoothed by using a kernel density estimator (KDE). The bandwidth used for the KDE is defined by the parameter ‘bw’. This allows setting bigger or smaller retention time deviations for peak groups. If the raw data is put into different subdirectories XCMS treats each subdirectory as different sample groups. An example may be classifying healthy patients as one sample group and ill patients as another one. To create a valid peak group a peak has to be found in a minimum number of samples of at least one sample group. This minimum number is defined by the parameter ‘minsamp’. A minimum fraction of samples of a sample group containing the same peak is defined by the parameter ‘minfrac’. The maximum number of peak groups to be found within each m/z slice is defined by ‘max’. Default values for all parameters of the ‘density’ method are listed in Table 3.

Table 3: Parameters and default values of the 'density' grouping method

Parameter	Default value
mzwid	0.25
bw	30
minfrac	0.5
minsamp	1
max	50

2.2.2.2. 'nearest'

Although this grouping method is part of XCMS it was inspired by mzMine (7). The parameters of the method and its respective default values are shown in Table 4. For every peak the nearest neighbors regarding m/z and retention time are calculated. This is done using the R-package RANN (31). Since retention time differences are normally much bigger than m/z differences they affect the distance between two peaks much more. To counter this problem a factor is defined by the parameter 'mzVsRTbalance'. The maximum number of neighbors can by limited by the parameter 'kNN'. Additionally peaks are only considered neighbors if they are within an m/z- and retention time range which can be defined the parameters 'mzCheck' and 'rtCheck' respectively. Finally all neighbors are combined to one peak group.

Table 4: Parameters and default values of the grouping method 'nearest'

Parameter	Default value
mzVsRTbalance	10
kNN	10
mzCheck	0.2
rtCheck	15

2.2.3. Retention time correction methods

Liquid chromatography does not always produce exactly the same retention times for each sample. To correct these chromatographic shifts retention time correction is necessary. Different methods for retention time correction exist. Some of them require a grouping step before the correction some of them don't.

2.2.3.1. 'obiwarp'

Grouping before retention time correction can be a problem. Ultra high performance liquid chromatography (UHPLC) can produce very sharp peaks. If substances with the same mass elute at a comparable time, shifts might cause the grouping to choose the wrong peak. To counter this problem, the retention time correction method 'obiwarp', which does not rely on grouping, was developed (32). Different parameters allow adjusting the method (Table 5). The method starts by choosing a sample as reference. By default the one containing the most detected peaks is chosen or the user may specify one using the parameter 'center'. First profiles are generated. The parameter 'profStep' defines the m/z width of these profiles. Then, by comparing these profiles a similarity matrix is calculated. The parameter 'distFunc' specifies the function to calculate distances between points in the profiles. Available distance functions are Pearson's R ('cor'), covariance ('cor'), product ('prd') and the Euclidean distance ('euc'). Per default the distance function 'cor_opt' is used which is an improved, faster version of 'cor'. Gap openings and gap enlargements are penalized by the parameters 'gapInit' and 'gapExtend' respectively. Each distance function has its own default value for 'gapInit' and 'gapExtend'. A gap is defined as a horizontal or vertical step in the two dimensional matrix. A weighting for diagonal moves and gap moves is specified by the parameters 'factorDiag' and 'factorGap' respectively. Using these parameters an additive score matrix is calculated. The parameter 'localAlignment' defines whether a local or a global gap penalty is given. In the score matrix an optimal path is found. Points on this path having a high score a kept as anchor points. Across this anchor points a smooth warp function is generated. The

parameter ‘response’ defines how many anchor points for the creation of the warp function are used with the value zero only using the start and end points.

Table 5: Parameters and default values of the retention time correction method ‘obiwarp’

Parameter	Default value				
center	NULL				
profStep	1				
response	1				
factorDiag	2				
factorGap	1				
localAlignment	0				
initPenalty	0				
distFunc	cor_opt	cor	cov	prd	euc
gapInit	0.3	0.3	0.1	0	0.9
gapExtend	2.4	2.4	1.7	7.8	1.8

2.2.3.2. ‘loess’

The retention time correction method ‘loess’ (locally weighted scatterplot smoothing) needs grouping before the correction can be done (4). It uses ‘well-behaved’ groups as temporary standards. These groups consist of peaks from almost all samples and only few samples contribute more than one peak, making a proper grouping very likely. How many peaks may be missing or how many peaks may be additional for a group to be still ‘well-behaved’ can be defined by the parameters ‘missing’ and ‘extra’ respectively (Table 6). For sections were no ‘well-behaved’ groups are present interpolation is needed. The method of interpolation is defined by the parameter ‘smooth’ and the degree of smoothing by the parameter ‘span’. Outlier removal is

provided by the parameter 'family' to gain robustness for initially wrong grouped peaks.

Table 6: Parameters and default values of the 'loess' retention time correction method

Parameter	Default value
missing	1
extra	1
smooth	c('loess', 'linear')
span	0.2
family	c('gaussian', 'symmetric')

2.2.4. Filling in data of missing peaks

In this processing step peaks from samples that are missing in a peak group are integrated. Therefore the median minimum retention time and median maximum retention time of the peaks within a group are defining the chromatographic width of the missing peaks. The same is done for the mass-to-charge ratio to get a range in the m/z axis. The created range in chromatographic and m/z axis is then integrated in the samples that are not contributing a peak to the peak group.

2.3. CAMERA

Mass spectrometry has to ionize substances to make them detectable. The ionization process can produce several ion species for each substance including fragment, isotopologue or different adduct ions. CAMERA is an R-package intended to reduce biological interpretation efforts by annotating these different ion species (33). Isotopologues are annotated using an m/z difference of $1.0033/z$[19]. Additionally the

intensity ratio is checked by defining the minimum number of carbon (C_{min}) as one and calculating the maximum number of carbon (C_{max}) by:

$$C_{max} = \frac{m/z^{monoiso}}{12} \quad (1)$$

By multiplying C_{min} and C_{max} with the natural abundance of ^{13}C (1.1%) a lower and an upper intensity ratio is defined.

2.4. IPO optimization approach

IPO is an R-package which uses isotopologue information as basis for one if its reliability scores (Isotopologue Parameter Optimization). The package is intended to optimize parameters of XCMS methods. All these parameters are optimized by applying a semi sequential approach (**Figure 5**). Therefore the parameters to be optimized are split. First, peak picking parameters are optimized separately and then the retention time correction and grouping parameters are optimized simultaneously. Splitting the parameters reduces necessary experiments and hence improves performance. Dividing the optimization into two separate steps can only be done with scores that are not influenced by the other step. This is hardly possible for retention time correction and grouping. Grouping is the process of combining peaks with similar m/z and retention times from different samples to metabolic features (peak groups). Deviations in the chromatographic dimension are emended by the retention time correction. For assessing the reliability of retention time correction the grouping step is needed. Due to this inherent connection IPO optimizes the parameters for retention time correction and grouping simultaneously. Experiments for testing different parameter levels are achieved by design of experiments (DoE). For every experiment of the DoE a score is calculated. These scores are evaluated using response surface models (RSM). In a ‘maximum focusing’ step the parameters giving the best score are found and the parameter levels adjusted accordingly. As long as the respective scores keep on increasing the optimization process continuous.

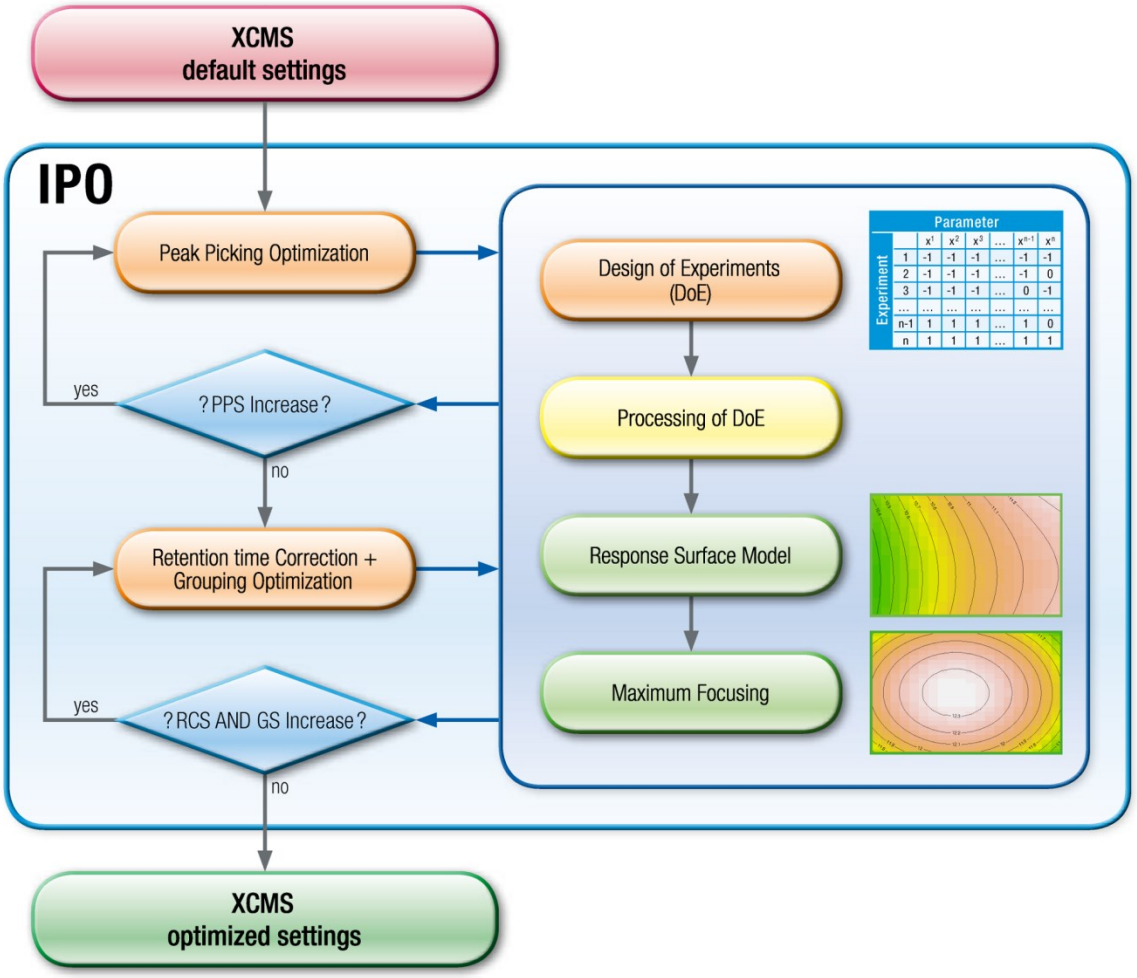


Figure 5: An illustration of the workflow for XCMS parameter optimization. XCMS default parameter levels are used at the start of the optimization process. First peak picking parameters are optimized, thereafter parameters for retention time correction and grouping. The optimization process is the same for both. The DoEs are created by using Central-Composite designs. The experiments of the DoE are processed and the respective scores calculated. These scores are evaluated by response surface models. In the maximum focusing the combination of parameters that yields the best score is found and the parameter levels are adjusted accordingly. The optimization process continues as long as the respective scores are increasing.

2.4.1. Design of Experiments

A designed experiment is a series of tests in which targeted modifications are made to the input variables of a process. DoE aims to either explain changes of the response variable or to optimize the response. Statistically a design of experiments refers to the process of planning an experiment in order to collect appropriate data for statistical analysis, resulting in valid and reliable conclusions (34). The Box-Behnken design (BBD)(35) and the center faced Central-Composite design (CCD) are both three level incomplete factorial designs for fitting a second order response surface model. In these designs three evenly distributed levels are set for each parameter. This implies that only quantitative parameters are supported. **Figure 6** shows an example of these levels for one parameter. The left circle represents the minimum level to be tested; the right circle shows the maximum level. In combination those two levels describe a parameter range. The circle in the middle defines a center point which is tested as well. These three levels are represented in a coded style by -1, 0 and 1. An example for the numerical levels these coded values represent is shown by the numerical series at the bottom of the figure.

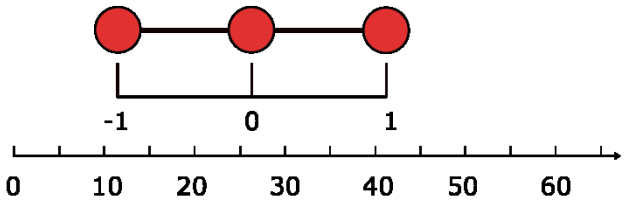


Figure 6: Illustration of three levels for one parameter. The values -1, 0 and 1 are the encoded values. The numerical series shows an example for the respective decoded values.

BBD needs $2k * (k - 1) + C_0$ experiments for k parameters and C_0 center points whereas the CCD needs $2^k + 2k + C_0$. This shows that BBD is slightly faster. This is achieved by laying less emphasis on testing extreme experiments on the lower and upper levels (36). However, CCD distributes the experiments among the three levels

equally, making it the better choice for parameter optimization. Therefore IPO-Versions from 1.5.6 use center faced CCD generated by the R-package rsm (37).

2.4.2. Peak picking optimization

The previously described peak picking methods ‘centWave’ and ‘matchedFilter’ are supported by IPO. The starting levels for optimization of the peak picking parameters are shown in Table 7. Parameters that are not in the table are only optimized by IPO if defined by the user. This can be done by defining a lower and an upper limit for the parameter. The parameters ‘mzCenterFun’, ‘integrate’, ‘fitgauss’ and ‘index’ are not supported by IPO since CCD cannot provide different test levels for qualitative parameters.

Table 7: IPO default levels for peak picking optimization

Parameter	Level 1	Level 2 (center point)	Level 3
findPeaks.centWave			
min peakwidth	10	20	30
max peakwidth	35	50	65
ppm	15	25	35
mzdiff	-0.001	0.0055	0.01
snthresh	10		
noise	0		
prefilter	c(3,100)		
mzCenterFun	‘wMean’		
integrate	1		
fitgauss	FALSE		

findPeaks.matchedFilter			
fwhm	25	30	25
snthresh	3	10	17
step	0.05	0.10	0.15
steps	1	2	3
sigma	fwhm / 2.3548		
max	5		
mzdiff	0.8 - step * steps		
index	FALSE		

For each experiment of the DoE first the peak picking is applied. Thereafter a novel peak picking score (*PPS*), which expresses the reliability of the peak picking result, is calculated according to the following formula:

$$PPS = \frac{RP^2}{\text{`\#all peaks'} - LIP} \qquad (2)$$

PPS is defined as the ratio of the squared number of *RP*s to the number of all peaks (‘*#all peaks*’) diminished by the number of *LIP*s. It relies on the existence of stable ^{13}C isotopes that naturally occur in all biological samples. Peaks belonging to an isotopologue are defined as reliable peaks (RP). IPO implements a novel method to identify isotopologues consisting of ^{13}C isotopic peaks that does not need any parameters set by the user. IPO characterizes isotope peaks by three criteria: The tolerable ranges of these criteria are defined by or can be calculated relative to the respective ^{12}C peak. The first criterion states that the isotope peak’s mass-to-charge ratio has to be within a certain mass range. Due to shifts of the mass spectrometry’s accuracy or measuring in profile mode every peak already spans a certain mass range. Increasing it by the difference of the exact mass of ^{13}C to ^{12}C creates a new range in which mass-to-charge ratio of the ^{13}C isotope peak must be found. Second, the isotope peaks must elute at the same time as the parent peaks. To restrict peaks on the retention time axis, a relative retention time window is specified. It states that a

^{13}C peak must be found within 0.5% of the ^{12}C-peak's median retention time. As a third criterion, the intensities of isotope peak candidates have to be within a certain intensity window. Based on a hydrocarbon chain (**Figure 7**) the maximum number of carbon atoms (*maxC*) is estimated by the following formula:

$$maxC = floor\left(\frac{m/z - 2*CH3}{CH2} + 2\right) \qquad (3)$$

m/z is the accurate mass-to-charge ratio (38) of a peak. *CH3* is the exact mass of a molecule consisting of one carbon atom and three hydrogen atoms whereas *CH2* defines a molecule with one carbon and two hydrogen atoms. With regard to the ends of a hydrocarbon chain m/z is diminished by two CH3 which is later corrected by '*+ 2*'. The remaining mass is divided by CH2. Eventually the function *floor* cuts of fractional digits giving the maximum possible number of carbons. The intensity window for a potential ^{13}C isotope (39) peak is calculated by assuming a minimum of one carbon atom and a maximum of maxC. Both values are then multiplied with the natural abundance of ^{13}C isotopes (1.109%) and the ^{12}C peak's intensity. Alternatively IPO (version > 1.6) allows to annotate isotope peaks using the R-package CAMERA (33).

Peaks with an isotopic peak concentration too low to measure are called ´low intensity peaks´ (*LIPs*). To calculate LIPs first all peak intensities which are not reliable peaks are ordered decreasing. Thereafter a threshold is defined by calculating the mean of the lower three percent of these intensities. Then, maxC is estimated for each peak. The estimated carbon atoms are then multiplied with the natural abundance of ^{13}C isotopes, yielding NAC. The peak`s intensity multiplied with its NAC is used for the determination of the maximum intensity of the natural ^{13}C isotope peak. If the intensity of the isotopic peak would be below the previously calculated threshold, the peak is considered as an LIP.

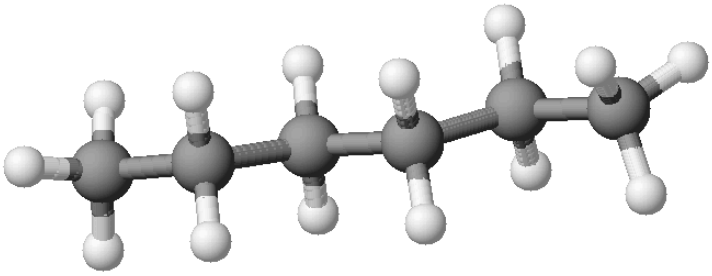


Figure 7: Example of a hydrocarbon chain. The black spheres represent carbon atoms, the white ones hydrogen. Carbon atoms can have four bonds. The carbon atoms at each end of the chain are bonded to three hydrogen atoms; the carbon atom in the middle can only be bonded to two hydrogen atoms.

2.4.3. Retention time correction and grouping optimization

The two processing steps after peak picking are retention time correction and grouping. Therefore an xcmsSet-object with the previously optimized settings from the peak picking optimization is created. But if only retention time correction and grouping parameters are to be optimized any xcmsSet-object can be used as well. This object is then used as basis for the optimization of these two processing steps. Similar to the optimization of the peak picking parameters a DoE with default starting values for is created (Table 8). Again, the qualitative parameters as ‘distFunc’, ‘plotType’, ‘smooth’, and ‘family’ are not supported by IPO. As default retention time correction method ‘obiwarp’ is used. For each experiment of the DoE a retention time correction and grouping target value (RGTV) is calculated. The calculation of this target value is presented subsequently. IPO (v1.7.0) supports the retention time correction method ‘loess’. In contrast to the ‘obiwarp’ method it needs grouping before it can be applied.

Table 8: IPO default levels for retention time correction method 'obiwarp' and grouping method 'density'

Parameter	Level 1	Level 2 (center point)	Level 3
retcor.obiwarp			
gapInit	0.0	0.2	0.4
gapExtend	2.1	2.4	2.7
profStep	0.7	0.85	1
distFunc	'cor_opt'		
plotType	'none'		
response	1		
factorDiag	2		
factorGap	1		
localAlignment	0		
initPenalty	0		
retcor.loess			
missing	1	2	3
extra	1	2	3
profStep	0.7	0.85	1
span	0.1	0.2	0.3
smooth	'loess'		
family	'gaussian'		
plotType	'none'		

group.density			
bw	22	30	38
minfrac	0.3	0.5	0.7
mzwid	0.015	0.02	0.035
minsamp	1		
max	50		

2.4.3.1. Grouping score (GS)

To create a grouping score the usage of pooled sample injections as basis for the optimization process is exploited. But any other kind of replicates or similar samples can be used as well. However, a pooled sample is best suited for the use in the optimization process since it contains all substances of a study population and is normally measured periodically within a study hence these samples also reflect drifts. Because the same sample was injected several times, peak groups where each injection of the sample contributes exactly one peak are defined as ´reliable groups´. All other peak groups are classified as ´non-reliable groups´. The ratio of the squared number of ´reliable groups´ to ´non-reliable groups´ is defined as the grouping score (GS):

$$GS = \frac{\#\text{`reliable groups'}^2}{\#\text{`non-reliable groups'}} \qquad (4)$$

2.4.3.2. Retention time correction score (RCS)

To increase the quality of retention time correction the chromatographic deviations within peak groups should be minimized. For every peak group a group retention time shift (GRTS) is calculated by:

$$GRTS(x) = \frac{\sum_{n=1}^{k} abs(median(x) - x_n)}{k} \quad (5)$$

The variable *x* represents all the retention times of the peaks of one peak group. The function *median* returns the median value of a set of values. The absolute value is calculated by the function *abs*. The variable *n* represents an index pointing to one specific element of a set of values; the variable *k* is the number of values within the set. GRTS is defined as the sum of all absolute differences of the retention times of one peak group to the peak group's median retention time divided by the number of retention times within the group. The average retention time shift (ARTS) is then calculated by taking the average of all GRTS values:

$$ARTS = \frac{\sum_{n=1}^{k} GRTS}{k} \quad (6)$$

ARTS has to be decreased to improve the quality of the retention time correction. The inverse of ARTS is calculated to create a score which gets better when increased by

$$RCS = \frac{1}{ARTS} \quad (7)$$

giving a retention time correction score (RCS).

2.4.3.3. Retention time correction and grouping target value (RGTV)

The combination of RCS and GS results in the retention time and grouping target value (RTGV) which is calculated as follows:

$$RGTV = norm(RCS) + norm(GS) \quad (8)$$

RCS and GS tend to be on different levels which would result in a different influence on RGTV. Simple multiplication does not work either since also the change of these two scores from one run to the next can be on complete different levels. Therefore the function *norm* is used which represents a unity based normalization step to achieve values in a range between zero and one. The function is applied on all RCS and GS

values of a DoE separately. Thereafter the normalized RCS and GS values of each experiment of one DoE are aggregated to create RGTV.

2.4.4. Response surface model

For every experiment of the DoE the respective score is calculated. The parameter levels defined in the DoE are called explanatory values and the score is the response. A response surface model (RSM) is then applied to estimate a full second-order model which describes the relation of explanatory values and response. The full second-order model is the combination of a ‘first-order’, a ‘two-way-interaction’ and a ‘pure quadratic’ model (37). If only one parameter is optimized a linear model is used. An example of different possibilities of RSM-presentations is given in **Figure 8**. The color corresponds to the response estimated at the respective levels. The contour plot although is clearer than the perspective plot since the third dimension of the perspective plot only shows the estimated response which can already be determined by the color.

a) **b)**

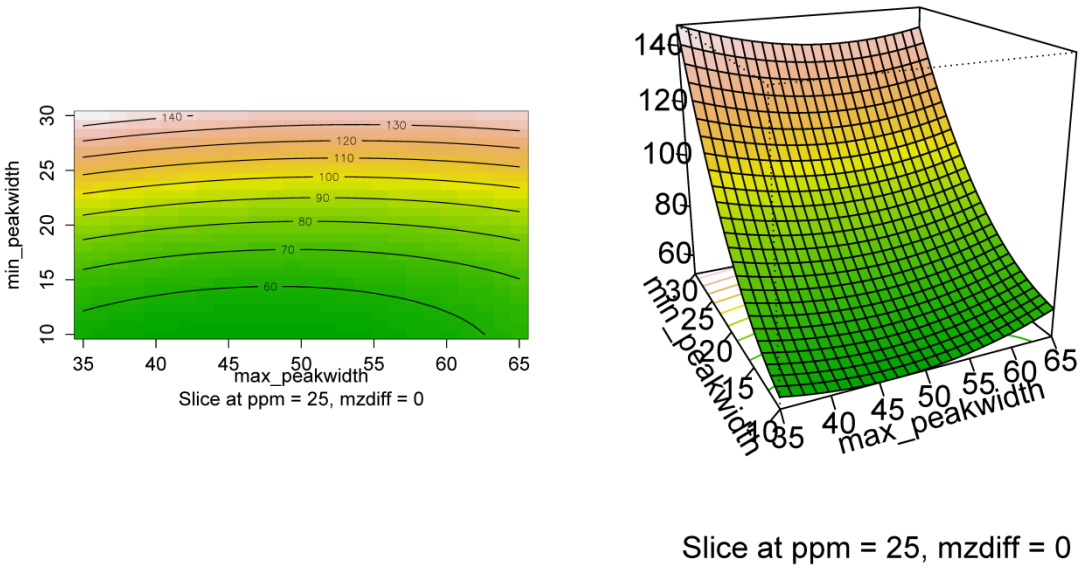


Figure 8: Examples of a response surface model. Both show the same RSM, a) as contour plot and b) as perspective plot. The colors represent the estimated responses to the different parameter levels.

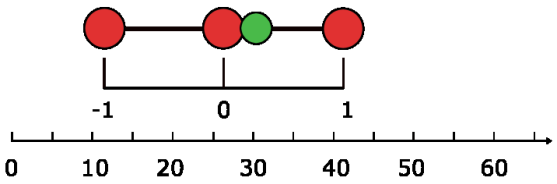
2.4.5. Maximum focusing

The ‘maximum focusing’ consists of three steps (**Figure 9**). Every parameter spans a certain range, resulting in *n* dimension for *n* parameter. First the RSM is used to find the parameter levels giving the highest response (PPS or RGTV) indicated by the green circle. This is done by creating a multidimensional matrix with different coded parameter levels between -1 and 1. The number of values between -1 and 1 depend on the number of dimensions because higher dimension would need too much memory and cause the optimization to stop. Therefore a step size of 0.05 is chosen for one dimension, 0.1 for up to five dimensions and 0.2 for more than five. This matrix and the response surface model are then used in the method ‘predict’ which estimates the response for all values in the matrix.

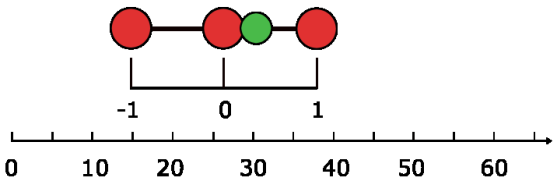
Defining the multidimensional matrix is of utmost importance as can be seen in **Figure 10** which shows contour plots of the parameters ‘min_peakwidth’ and

'max_peakwidth' cut at different slices. **Figure 10**a shows the contour of the slice at the lower levels for the parameters 'ppm' and 'mzdiff', **Figure 10**b and c show the contour of the slice at the center points and upper levels respectively and **Figure 10**d show the contours of the slice at the best levels for this DoE. It can be seen that for each contour the response greatly differs. Also, the highest response regarding the parameters 'min_peakwidth' and 'max_peakwidth' can be found at different levels in the four illustrated plots.

a)



b)



c)

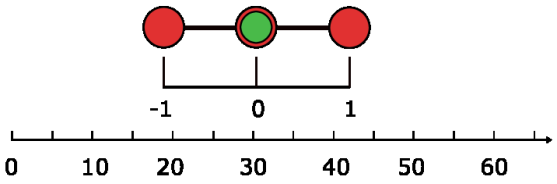


Figure 9: This illustrates the three step of 'maximum focusing'. In figure a) the level achieving the maximum respond is found (green circle) within a deviation of 25% of the center point. Figure b) illustrates the decrease of the level range and in c) the center point is shifted to the level of the maximum response.

The second step is to adjust the parameter ranges (**Figure 9**b). If the best level for one parameter is within a deviation of 25% of its center point, the range for the next DoE is decreased by 20% (‘zooming in’). If both, the minimum and maximum levels achieve the best response, the range is increased by 20% (‘zooming out’). For some parameters minimum and maximum levels are defined in IPO. The parameter ‘mzdiff’ can be negative for the ‘centWave’ peak picking method and has a defined minimum of 0.001 for the ‘matchedFilter’ method (Table 9). Also, the ‘max_peakwidth’ must never be greater than ‘min_peakwidth’. For all other peak picking parameter the minimum is set to one, for the retention time correction and grouping parameters the minimum is zero. A 20% ‘zooming in’ is done if the best level is found at an absolute minimum of a parameter.

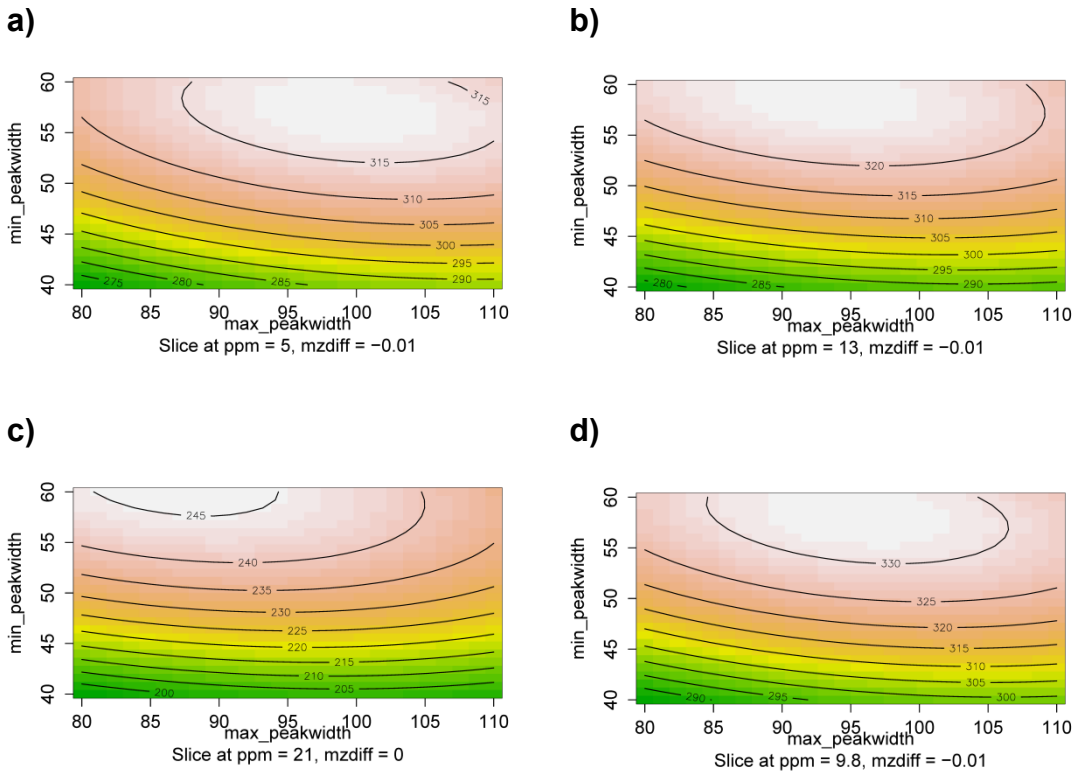


Figure 10: Four response surface models cut at different parameter levels: a) at minimum level; b) at the center point; c) at maximum level; d) at best settings

Table 9: Lower and upper limits of parameters defined by IPO

Parameter	Lower limit	Upper limit
min peakwidth	3	-
max peakwidth	min peakwidth	-
mzdiff	-100,000,000 0.001	-
step	0.0005	-
span	0..001	-
profStep	0.3	1
bw	0.25	-
minFrac	-	1
mzwid	0.0001	-

In the third step of 'maximum focusing' the levels producing the highest response are used as new center points for the next DoE illustrated in **Figure 9**c. The thus newly generated DoE is then calculated again. As long as the respective scores increase, this process continues.

2.5. Computational platform

LC-MS raw files originating from the central carbon metabolism dataset were converted using 'msConvert' of the ProteoWizard release 3.0.5033 (40,41). The files from the other datasets were converted to mzXML using 'ReadW' 4.0.2. For the parameter optimization a server and a personal computer were used. The server was an Intel(R) Xeon(R) CPU E5-2620 @ 2.00GHz system with 64 GB RAM running Debian GNU/Linux 7.5 (v3.2.0-4-amd64) with R (v3.0.0) and using the R-packages XCMS (v.1.42.0), Rmpi (v.0.6-5) as well as rsm (v.2.07). The personal computer was an Intel(R) Core™ i5 CPU 760 @ 2.80GHz system with 4 GB RAM running Windows

7 32 Bit with R (v3.1.1) and using the R-packages XCMS (v.1.40.0), rsm (v.2.07). The installed IPO version is described within the methods of the respective datasets.

2.6. Datasets

Dataset from different sample types, sample processing methods, liquid chromatography and mass spectrometry were used to evaluate IPO.

2.6.1. HILIC dataset - CARDIONOR

The CARDIONOR study (42,43) aimed to identify type-2 diabetic patients that are not responding to intensive, multifactorial treatment for atherosclerosis (44–47). A total of 106 serum samples (one sample per test subject at baseline before the start of intensified T2D-therapy) were processed using a acetone extraction method (48). From the 106 extracts, a quality control (pooled sample or QC) was mixed taking an aliquot of each sample. This QC sample was analyzed periodically after every third sample. Additionally, a blank sample, for subsequent use in filters, has been measured before each QC-sample (**Figure 11**). The samples were split into three aliquots and stored for additional measurements at -80 °C. Due to blank and pooled QC samples as well as due to the splitting into five batches a total of 180 samples had to be measured. LC-MS analysis was done with a Thermo Exactive mass spectrometer. Chromatographic separation was achieved using hydrophilic interaction chromatography (HILIC). HILIC was performed on a Luna NH2 column (2×150 mm; 3 µm; Phenomenex, Torrance, USA) following Bajad et al. (49). Full scan spectra were recorded in positive-negative switching electrospray from m/z 85–1,700 with a resolution of 25,000 (m/z 200). For the optimization process all quality control injection were used. The 46^{th} sample was removed from the dataset because of problems within the LC-HRMS run. All other raw files were converted using ReadW (v.4.0.2) and split by ionization mode into separate files using in-house software. The optimization was done on the server using IPO (v1.5.6) and started with default values for all parameters except for ‘noise’ which was set to a fixed level of 1,000.

001_Bl	019_Sa	037_QC	055_Bl	073_Sa	091_QC	109_Sa	127_Sa	145_Bl	163_Sa
002_QC	020_Sa	038_Sa	056_QC	074_Sa	092_Sa	110_QC	128_Sa	146_QC	164_Sa
003_Sa	021_Bl	039_Sa	057_Sa	075_Bl	093_Sa	111_Bl	129_Sa	147_Sa	165_Sa
004_Sa	022_QC	040_Bl	058_Sa	076_QC	094_Sa	112_Sa	130_Bl	148_Sa	166_Bl
005_Sa	023_Sa	041_QC	059_Sa	077_Sa	095_Bl	113_Sa	131_QC	149_Sa	167_QC
006_Bl	024_Sa	042_Sa	060_Bl	078_Sa	096_QC	114_Sa	132_Sa	150_Sa	168_Sa
007_QC	025_Sa	043_Sa	061_QC	079_Sa	097_Sa	115_Bl	133_Sa	151_Bl	169_Sa
008_Sa	026_Bl	044_Sa	062_Sa	080_Bl	098_Sa	116_QC	134_Sa	152_QC	170_Sa
009_Sa	027_QC	045_Bl	063_Sa	081_QC	099_Sa	117_Sa	135_Bl	153_Sa	171_Bl
010_Sa	028_Sa	046_QC	064_Sa	082_Sa	100_Bl	118_Sa	136_QC	154_Sa	172_QC
011_Bl	029_Sa	047_Sa	065_Bl	083_Sa	101_QC	119_Sa	137_Sa	155_Sa	173_Sa
012_QC	030_Sa	048_Sa	066_QC	084_Sa	102_Sa	120_Bl	138_Sa	156_Bl	174_Sa
013_Sa	031_Bl	049_Sa	067_Sa	085_Bl	103_Sa	121_QC	139_Sa	157_QC	175_Bl
014_Sa	032_QC	050_Bl	068_Sa	086_QC	104_Sa	122_Sa	140_Bl	158_Sa	176_QC
015_Sa	033_Sa	051_QC	069_Sa	087_Sa	105_Bl	123_Sa	141_QC	159_Sa	177_Sa
016_Bl	034_Sa	052_Sa	070_Bl	088_Sa	106_QC	124_Sa	142_Sa	160_Sa	178_Sa
017_QC	035_Sa	053_Sa	071_QC	089_Sa	107_Sa	125_Bl	143_Sa	161_Bl	179_Bl
018_Sa	036_Bl	054_Sa	072_Sa	090_Bl	108_Sa	126_QC	144_Sa	162_QC	180_QC

Figure 11: Overview of the samples measured within the CARDIONOR study. The number shows the order in which the samples have been measured and the text shows the type of sample. ‘Bl’ denote blank samples and ‘Sa’ is the code for serum samples. The quality control sample injections are represented by the tag ‘QC’. The quality control with the red background was removed from the dataset.

2.6.1.1. Optimization confidence tests

All 36 measurements of the pooled sample in positive ionization mode were used in a setup to determine the confidence and stability of the parameter optimization using IPO. A different number of optimizations were calculated using different sample sets following:

- 64 optimizations using 2 measurements
- 32 optimizations using 4 measurements
- 16 optimizations using 8 measurements
- 8 optimizations using 16 measurements
- 4 optimizations using 32 measurements

The samples for the sets were chosen randomly and no set consisted of the same measurements. The optimizations were done using IPO v1.6 and the parameter 'checkBorderIntensity' was set to true. The confidence tests optimized the 'centWave' peak picking parameters 'min_peakwidth', 'max_peakwidth', 'ppm' the retention time correction parameters 'gapInit' and 'gapExtend' and the grouping parameters 'bw' and 'mzwid'. The starting levels for 'min_peakwidth' were set to 25, 35 and 45, the levels for 'max_peakwidth' to 55, 70 and 85 and the levels for ppm to 10, 15 and 20. The parameter 'noise' was fixed at 1,000, 'profStep' at 1 and 'minfrac' at 0.75. All other parameters were kept at their default levels.

2.6.2. HILIC dataset – Bariatric Surgery

44 patients underwent bariatric surgery (50–52). Serum samples were taken before the surgery, right after the surgery and at a 6 month follow up resulting in 132 serum samples. These samples were processed as published by Yuan et al (53). From all samples, 10 µl were taken and mixed together to generate a pooled sample. LC-MS analyses were performed with an Ultimate 3000 UHPLC system (Thermo Fisher Scientific, San Jose, CA, USA) coupled to a high resolution mass spectrometer Q-Exactive (Thermo Fisher Scientific, Bremen, Germany). Chromatographic separation was achieved on a Luna NH2 column (2×150 mm; 3 µm; Phenomenex, Torrance, USA) following the procedure published by Bajad et al (49). Full scan spectra were recorded in switching ionization mode from m/z 70–1,050 with a resolution of 140,000 (m/z 200). After every third sample a pair of a blank and a pooled sample were measured. To achieve an unbiased evaluation of the optimization approach measurements were separated into a training set and a test set. The training set consisted of twelve measurements of the pooled sample and the test set was made up of eleven different injections of the pooled sample. The optimization was done using IPO v1.6 on the server. The peak picking parameters 'min_peakwidth', 'max_peakwidth', 'ppm' and 'mzdiff' were optimized. Isotopic peaks were identified setting the parameter 'checkBorderIntensity' to true. For the retention time correction and grouping parameter optimization the parameters 'profStep', 'gapInit', 'gapExtend',

'minfrac' and 'mzwid' were chosen. The parameter 'bw' was set to 15 and all other parameters were kept at XCMS default values.

2.6.3. RP-HPLC method - Lipidomics

Tissue samples from muscle, lung and proventriculus of mice were collected and processed, using the method published by Fauland et al (54). LC-MS analyses were performed with an Ultimate 3000 HPLC system (Thermo Fisher Scientific, San Jose, CA, USA) coupled to a high resolution mass spectrometer Q-Exactive (Thermo Fisher Scientific, Bremen, Germany). Chromatographic separation was achieved on a Hypersil GOLD column (100 mmx1 mm, 1.9 µm; Thermo Fisher Scientific, San Jose, CA, USA). Full scan spectra were recorded from m/z 350–1,050 with a resolution of 140,000 (@ m/z 200) using heated positive electrospray ionization. After every sixth sample a pair of a blank and a pooled sample was measured. The eight measurements of the pooled sample injections were split into a training set and a test set consisting of four measurements each. On the personal computer IPO (v1.5.6) was used to optimize the XCMS parameters.

2.6.4. IP-RP-HPLC method - Central carbon metabolism

Metabolites were extracted from stationary phase yeast cells (BY4741 background strain). Culture aliquots of OD_{600} were harvested by filtration with 0.22 µm sterile filters, washed once (on filter) with 5 ml double-distilled water and were immediately quenched by deep-freezing the filters in liquid nitrogen. The filtration and the washing step were performed in less than 30 seconds before the freezing step. For the acid extraction of metabolites, cells were resuspended in 1 ml ice-cold 5% trichloroacetic acid (TCA) and incubated for 1 hour on ice with occasional vortexing. Supernatants (10 min; 10,000 g) were lyophilized and resuspended in 200 µl double-distilled water. Aliquots of each cell extract were pooled.

LC-MS analyses were performed with an Ultimate 3000 HPLC system (Thermo Fisher Scientific, San Jose, CA, USA) coupled to a high resolution mass spectrometer

Exactive™ Orbitrap system used in profile mode. Chromatographic separation was achieved on an Atlantis T3 C18 analytical column (150 mm x 3 mm, 3 µm, Waters, USA). HPLC was run with a two eluent multi-step gradient of 2-propanol and an aqueous mobile phase (5 % methanol (*v/v*), 10 mM tributylamine (TBA) and 15 mM acetic acid, pH 4.95) within 40 minutes per sample (55).

Heated electrospray ionization was used for negative ionization. Data acquisition was conducted via full scan of all masses between 70 and 1,100 m/z (R = 50,000). The injection volume was set to 10 µl per sample. A blank sample and a pooled sample were measured periodically after every fifth sample. The measurements were done in two separate experiments (batches). The pooled sample injections of the first batch were used as training set, the ones from the second batch as test set. On the server IPO (v1.6.2) was used on 16 cores to optimize the parameters ‘fwhm, ‘step’, ‘steps, ‘profStep’, ‘gapInit’, ‘gapExtend’, ‘minfrac’ and ‘mzwid’. The starting values of the parameter ‘step’ were set to 0.005 and 0.015. The parameter ‘mzdiff’ was set to 0.0075, ‘snthresh’ and to 10. This is the only dataset which used the peak picking method ‘matchedFilter’.

3. Results

Within this thesis a new approach for parameter optimization was developed and resulted in a freely available R-package called IPO. IPO was tested on several different datasets steaming from different sample types and measured with different LC-methods on different mass spectrometry devices for an in depth evaluation of IPO's reliability.

3.1. IPO – R-package and publication

The methods and innovations described in chapter '2.4 - IPO optimization approach' were combined into an R-package. The R-package IPO is freely available from https://github.com/glibiseller/IPO. IPO supports the XCMS peak picking methods 'matchedFilter' and 'centWave', the retention time correction method 'obiwarp' and 'loess' as well as the grouping method 'density'. IPO was published as an open access article in the scientific journal BMC Bioinformatics which is a top ten journal in the category MATHEMATICAL & COMPUTATIONAL BIOLOGY (56). The Appendix includes the published article as well as posters regarding parameter optimization which were presented at different conferences in Washington DC, Glasgow, London, Graz and San Francisco. Additionally to the source code an R-package contains help files and a vignette. A vignette consists of installation information for the package as well as described use case. These files are included in the Appendix as well.

3.2. Datasets

IPO was used on the datasets presented in section '2.6 - Datasets'. Comparison of the data achieved with default settings and the data calculated with the optimized settings for each dataset is presented.

3.2.1. HILIC dataset – CARDIONOR

This dataset consisted of a total of 180 samples. One pooled sample measurement was dismissed due to a problem within the LC-HRMS run. The remaining 36 pooled sample injections were used for optimization of peak picking parameters. Since retention time shifts may occur from sample to sample, the retention time correction and grouping parameters were optimized using all measurements except the blank samples (36 pooled sample injections + 106 serum sample injections). The blank samples should not contain any reliable data and were therefore not used within the optimization process.

3.2.1.1. Peak picking parameter optimization

The optimized peak width settings for the data achieved by negative electro spray ionization (nESI) had a minimum of 73.2 seconds and a maximum of 117.0 seconds as well as 69.0 and 97.8 seconds for positive electro spray ionization (pESI) respectively (Table 10). For nESI the 'ppm' parameter was with 30.0 slightly higher than the default value, for pESI with 25 the same as the default settings. The parameter 'mzdiff' achieved with nESI an optimized value of -0.00936 and with pESI -0.0153.

Table 10: HILIC-CARDIONOR - Optimized peak picking parameter

ESI mode	min peak width	max peak width	ppm	mzdiff
negative	73.2	117.0	30.0	-0.00936
positive	69.0	97.8	25.0	-0.0153

Using these optimized values on all 179 samples for the calculation of xcmsSet-objects for nESI and pESI lead to an increase of PPS from 5,084 to 15,263 (+200%) for nESI and from 3,986 to 14,947 (+278%) for pESI (Table 11). The number of peaks

was reduced from 1,136,832 to 712,075 whereas the number of reliable peaks increased from 64,284 to 80,217. An even bigger reduction of peaks was achieved for pESI where the number decreased from 1,360,975 to 729,487 whereas the reliable peaks were increased from 66,230 to 80,119 for pESI.

Table 11: HILIC-CARDIONOR - Comparison of peak picking result achieved with default and optimized settings

	Default negative ESI	Optimized negative ESI	Default positive ESI	Optimized positive ESI
PPS	5,084	15,263	3,958	14,947
# peaks	1,136,832	712,075	1,360,975	729,487
# reliable peaks	64,284	80,217	66,230	80,119

The optimization for nESI finished after 8 DoEs and 1.5 days, the one for pESI after 6 DoEs and 1.4 days using 16 cores on the server.

3.2.1.2. Retention time correction and grouping parameter optimization

The optimized values for the parameter ‘gapInit’ were with 0.59360 for nESI and 0.72064 for pESI higher than the default value (Table 12). Also the parameter ‘gapExtend’ registered a small increase from 2.4 to 2.42592 for nESI and 2.72592 for pESI. For both ionization methods the optimized levels for the parameter ‘profStep’ were with 0.884 for nESI and 0.652 for pESI lower than the default setting of 1. The optimized settings for the parameter ‘bw’ lay with 0.88 for nESI and pESI far below the default value of 30. The parameter ‘minfrac’ increased from the default 0.5 to 1.0 for nESI and pESI. The achieved values for the ‘mzwid’ parameter were 0.00732 for nESI and for pESI. The optimization on the server using 16 cores and 142 measurements finished for both ionization methods after four DoEs which both took about four hours to calculate.

Table 12: HILIC-CARDIONOR - Optimized parameters for retention time correction and grouping

ESI mode	gapInit	gapExtend	profStep	bw	minfrac	mzwid
negative	0.59360	2.42592	0.884	0.88	1	0.00732
positive	0.72064	2.72592	0.652	0.88	1	0.00732

Using the optimized setting on the whole dataset without filling missing peaks led to a RCS of 251.9 for nESI and 152.2 for pESI (Table 13). The values achieved with default settings were much lower for nESI and pESI with 23.9 and 17.1 respectively. The number of 'reliable groups' increased from 2 to 12 for nESI and from 1 to 66 for pESI. At the same time the number of 'non-reliable groups' was reduced from 3,314 to 1,575 for nESI and from 3,835 to 2,252 for pESI. GS was very low for nESI and pESI, default as well as optimized.

Table 13: HILIC-CARDIONOR - Comparison of result achieved after retention time correction and grouping with default and optimized settings

	Default negative ESI	Optimized negative ESI	Default positive ESI	Optimized positive ESI
RCS	23.9	251.9	17.1	152.2
# reliable groups	2	12	1	66
# non-rel. groups	3,314	1,575	3,835	2,252
GS	0.00	0.09	0.00	1.93

3.2.1.3. Optimization confidence test

The 124 optimizations of the confidence tests finished after 45 days using 8 cores of the server. R's summary of the optimized parameters for the confidence tests are shown in Table 14. It can be seen that some minima and maxima are far from the

expected values for HILIC. But median, mean as well as 1st and 3rd quantile are well within expected ranges for all parameters except 'bw' where the 1st quantile is only 0.25.

Table 14: Output of summary for the parameter levels achieved from the optimizations of the confidence test

	min PW	max PW	ppm	gapInit	gapExtend	bw	mzwid
Min.	65.0	86.0	17.0	0.000	1.150	0.25	0.003
1st Qu.	74.0	145.0	19.9	0.120	2.100	0.25	0.015
Median	89.8	157.0	21.4	0.220	2.400	12.4	0.038
Mean	87.6	150.6	22.3	0.252	2.257	9.20	0.037
3rd Qu.	98.5	160.0	24.8	0.340	2.400	12.4	0.047
Max.	109.0	208.5	32.4	1.170	3.350	22.0	0.115

3.2.2. HILIC dataset – Bariatric Surgery

For the optimization a training set containing LC-HRMS data of twelve injections of a pooled sample were used. The achieved parameters were used on a test set consisting of eleven measurements of different injections of the same pooled sample.

3.2.2.1. Peak picking parameter optimization

The optimization resulted in a minimum peak width of 45.6 seconds and a maximum peak width of 99.5 seconds for nESI. For pESI the optimized minimum peak width was 42.4 seconds and the maximum peak width 95.0 seconds. For nESI the optimized 'ppm' parameter value was 12.25 and for pESI 11.75. The optimized value for the parameter 'mzdiff' was 0.0034 for negative ESI and 0.0021 for positive ESI (Table 15).

Table 15: HILIC-Bariatric - Optimized peak picking parameter

ESI mode	min peak width	max peak width	ppm	mzdiff
negative	45.6	99.5	12.25	0.0034
positive	42.4	95.0	11.75	0.0021

Comparing the data achieved with default and optimized settings on the test set PPS increased more than tenfold, from 29 to 392 for negative ESI and doubled, from 458 to 940 for positive ESI. The total number of peaks decreased from 50,990 to 39,087 in nESI and from 65,851 to 39,206 in pESI. Due to the optimization the reliable peaks increased from 1,165 to 3,505 in nESI and from 5,112 to 5,439 in pESI (Table 16).

Table 16: HILIC-Bariatric - Comparison of peak picking result achieved with default and optimized settings

	Default negative ESI	Optimized negative ESI	Default positive ESI	Optimized positive ESI
PPS	29	392	458	940
# peaks	50,990	39,087	65,851	39,206
# reliable peaks	1,165	3,505	5,112	5,439

The optimization of the peak picking parameters using eight cores on the server finished after five DoEs and 14 hours for nESI and after four DoEs in 11 hours for pESI.

3.2.2.2. Retention time correction and grouping parameter optimization

The optimization process on the server using the training set of nESI the values 0.64, 2.64, 0.73, 12.4, 0.94 and 0.003 for the paramters ‘gapInit’, ‘gapExtend’, ‘profStep’,

'bw', 'minfrac' and 'mzwid' respectively. For the training set of pESI the settings 0.64, 2.4, 0.87, 12.4, 0.94 and 0.003 were achieved (Table 17).

Table 17: HILIC-Bariatric - Optimized parameters for retention time correction and grouping

ESI mode	gapInit	gapExtend	profStep	bw	minfrac	mzwid
negative	0.64	2.64	0.73	12.4	0.94	0.003
positive	0.64	2.40	0.87	12.4	0.94	0.003

Comparing the default and optimized settings used on the eleven injections of a pooled sample of the nESI test set RCS increased fivefold, from 39.9 to 211.0. The reliable groups increased from 159 to 1,466 and the 'non-reliable groups' decreased from 1,349 to only 26. The comparison of the default and optimized settings on the pESI test set showed an increase of RCS from 9.7 to 120.9. The reliable groups increased from 314 to 857 and the 'non-reliable groups' decreased from 1,810 to 62 (Table 18).

Table 18: HILIC-Bariatric - Comparison of result achieved after retention time correction and grouping with default and optimized settings

	Default negative ESI	Optimized negative ESI	Default positive ESI	Optimized positive ESI
RCS	39.9	211.0	9.7	120.9
# reliable groups	159	1,466	314	857
# non-rel. groups	1,349	26	1,810	62
GS	19	82,660	54	11,846

The optimization of the retention time correction and grouping parameters finished after three DoEs and 48 minutes for nESI and after three DoEs and one hour for pESI -on the server using eight cores.

3.2.3. RP-HPLC method - Lipidomics

The dataset contained eight measurements of injections of a pooled sample. Four of these measurements were used as a training set and the remaining four as test set. The lipidomics method was performed only in positive ionization mode.

3.2.3.1. Peak picking parameter optimization

The optimization done on the training set proposed a minimum peak width of 26.8 and a maximum peakwidth of 69.0. The optimized 'ppm' parameter lay at 7.0 and the optimized 'mzdiff' parameter at -0.0065 (Table 19).

Table 19: RP_HPLC-Lipidomics - Optimized peak picking parameter

ESI mode	min peak width	max peak width	ppm	mzdiff
positive	26.8	69.0	7.0	-0.0065

Comparing the result of the optimized and default parameter settings showed a PPS increase of 59% from 9,001 to 14.283. The number of peaks decreased from 34,415 to 33,192 and the reliable peaks increased from 12,999 to 14,817 comparing default and optimized settings (Table 20).

Table 20: RP_HPLC-Lipidomics - Comparison of peak picking result achieved with default and optimized settings

	Default positive ESI	Optimized positive ESI
PPS	9,001	14,283
# peaks	34,415	33,192
# reliable peaks	12,999	14,817

The optimization process of the peak picking parameters finished after three DoEs and four hours using four cores of the personal computer.

3.2.3.2. Retention time correction and grouping parameter optimization

The optimization of the retention time correction parameter resulted in 0.4288, 2.1912 and 0.604 for the parameters ‘gapInit’, ‘gapExtend’ and ‘profStep’ respectively. The parameter settings achieved for grouping were 0.25, 1.0 and 0.00538 for the parameter ‘bw’, ‘minfrac’ and ‘mzwid’ (Table 21). For evaluation the default and optimized settings were used on the test set.

Table 21: RP_HPLC-Lipidomics - Optimized parameters for retention time correction and grouping

ESI mode	gapInit	gapExtend	profStep	bw	minfrac	mzwid
positive	0.4288	2.1912	0.604	0.25	1.0	0.00538

Comparing the default and optimized settings RCS increased from 37.6 to 602.0. The number of ‘reliable groups’ increased from 1,564 to 5,865 and the number of ‘non-reliable groups’ dropped from 3,248 to 35 (Table 22). The optimization of the retention time correction and grouping parameter finished after 5 DoEs and took 52 minutes.

Table 22: RP_HPLC-Lipidomics - Comparison of result achieved after retention time correction and grouping with default and optimized settings

	Default positive ESI	Optimized positive ESI
RCS	37.6	602.0
# reliable groups	1,564	5,865
# non-rel. groups	3,248	35
GS	753	982,806

3.2.4. IP-RP-HPLC method - Central carbon metabolism

The central metabolism dataset was optimized on the server using six measurements of injections of a pooled sample. The test set consisted of the same amount and kind of samples but were processed and measured in a second batch. The used LC-HRMS method only utilizes negative ionization. The optimization was performed on the personal computer using four cores.

3.2.4.1. Peak picking parameter optimization

The optimization of the peak picking parameter needed six DoEs and 6.8 hours to finish. The full width at half maximum (fwhm) parameter achieved by the optimization was 44.0. Furthermore the optimization process resulted in a ‘step’ parameter setting of 0.0061, ‘steps’ setting of 3 and a ‘mzdiff’ setting of 0.0075 (Table 23).

Table 23: IP-RP-HPLC - Optimized parameters for retention time correction and grouping

ESI mode	fwhm	step	steps	mzdiff
negative	44.0	0.0061	3	0.0075

A comparison of the data achieved from using default and optimized settings on the test set showed an increase of PPS by 97% from 59 to 116 (Table 24). The number of peaks increased from 1,405 to 1,445 and the number of reliable peaks increased from 200 to 278 (+39%).

Table 24: IP-RP-HPLC - Comparison of result achieved after retention time correction and grouping with default and optimized settings

	Default negative ESI	Optimized negative ESI
PPS	59	116
# peaks	1,405	1,445
# reliable peaks	200	278

3.2.4.2. Retention time correction and grouping parameter optimization

After three DoEs which finished after seven minutes the optimization resulted in the settings 0.592, 2.4, 1.0, 0.94 and 0.003 for the parameters ‘gapInit’, ‘gapExtend’, ‘profStep’, ‘minfrac’ and ‘mzwid’ (Table 25).

Table 25: IP-RP-HPLC - Default and optimized parameters for retention time correction and grouping

ESI mode	gapInit	gapExtend	profStep	minfrac	mzwid
negative	0.592	2.3	1.0	0.94	0.003

The optimized and the default settings were used on the test set. The comparison of these showed increase of RCS from 79.5 to 257.0 (+223%). Furthermore the reliable

groups increased by 27% from 150 to 190 whereas all 69 ‘non-reliable groups’ were eliminated comparing the result of the default and optimized settings (Table 26). This led to an 100-fold increased GS.

Table 26: IP-RP-HPLC - Comparison of result achieved after retention time correction and grouping with default and optimized settings without fillPeaks

	Default negative ESI	**Optimized negative ESI**
RCS	79.5	257.0
# reliable groups	150	190
# non-rel. groups	69	0
GS	326	36,481

4. Discussion

This chapter compares IPO and previously published optimization approaches as well as the various improvements implemented in IPO and problems solved are being discussed. Additionally the differences of the data achieved with default XCMS settings against data generated with optimized settings are focused.

4.1. IPO versus previous optimization approaches

IPO is intended to use replicate samples or repeated measurements of a pooled sample. Using a pooled sample is beneficial because it contains all substances from all samples mixed together. Therefore, all these substances are integrated in the optimization process. These pooled samples are usually used for normalization and drift correction and are therefore injected periodically throughout all measurements of a study. Therefore, information from the whole study including potential sensitivity losses, mass shifts or retention time drifts is considered in the optimization process. A DoE approach for parameter optimization which uses a series of diluted samples has already been described earlier (19). Eliasson et al. defined peaks as reliable ones which correlated with the dilution series of a pooled sample and calculated a reliability index. Because of the sensitivity losses throughout a whole study, periodical measurements of diluted samples are not reliable, since the intensity will not correlate with the dilution series anymore. However, a dilution series measured at the beginning does not contain drifts and sensitivity losses. Another disadvantage arises from the usage of a dilution series itself. Substances with low concentration may not be detectable anymore after dilution. Therefore these substances will not correlate with the dilution series and may even be classified as unreliable peaks. The approach published by Eliasson et al needed rather long to finish when optimizing all parameters of peak picking, retention time correction and grouping within one DoE which they called global approach. To decrease the optimization time Zheng et al (20) introduced a screening step before the actual optimization using a Plakett-Burman design. The screening step is used to determine XCMS parameters with high

influence on the reliability index. The design is a two level fractional factorial design which allows accomplishing a screening step with relatively few experiments. The two levels denote that two different values are tested for each XCMS parameter. Then, only those parameters which had a significant, positive influence on the reliability index were optimized after the screening step. However, testing only two values to determine the significance of the parameter's influence may lead to suboptimal results. An exemplary response to different parameter settings is shown in **Figure 12**. It can be seen that the biggest changes in the response occur between the parameter settings 15 and 27. However, if only the values 30 and 40 were tested, the parameter would be dismissed because no significant influence on the response would be shown. Therefore, these kind of screening experiments which only test two levels for each parameter always hold some rest risk. Also for three level designs the levels have to be chosen carefully. Testing the levels 10, 30 and 50 would not yield significant changes in the response; as for testing the levels 10, 20 and 30 the change would be highly significant. It can be concluded that well-chosen minimum and maximum levels (parameter ranges) are important.

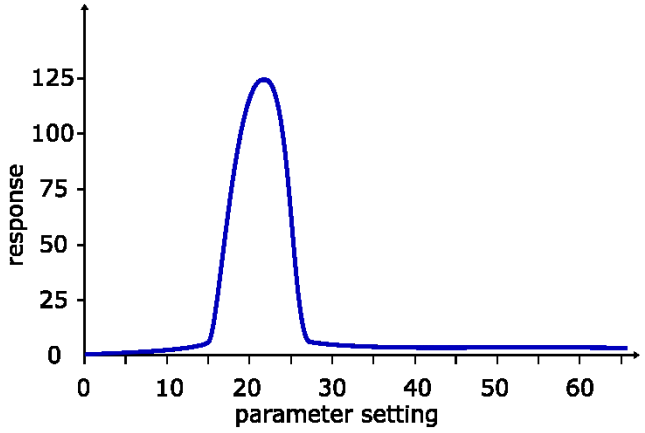


Figure 12: This figure shows an exemplary response which would be achieved by different parameter settings.

Another problem of the approach published by Zheng et al was that only parameters which had a positive influence on the reliability index were optimized. This was insufficient because also parameters with negative, significant influence need optimization.

To reduce calculation time Eliasson et al simulated another approach where peak picking, retention time correction and grouping parameters were sequentially optimized. But to calculate the reliability index based on the dilution series, retention time correction and grouping were needed. This was not a problem in the global approach since all processing steps and their respective parameters were included in the DoE. However, in the sequential approach the respective following steps in the data processing were performed with default settings. In contrast to the reliability index based on the dilution series the PPS introduced in this thesis can be evaluated for each sample individually. Therefore default parameters are not needed for retention time correction and grouping and the additional time needed for these two steps can be saved. After the creation of the DoEs, IPO calculates the experiments in parallel to diminish the optimization time. XCMS does support parallel processing but only for the peak picking step. With our approach also retention time correction and grouping are parallelized which again saves optimization time.

IPO's optimization process takes the default parameter settings of XCMS as center values. Therefore, the tool is also well applicable for inexperienced users. In addition to the center of a parameter, the parameter range used is highly significant as discussed previously (**Figure 12**). Therefore the parameter ranges are continuously adapted during the optimization process in the 'maximum focusing' step as described in the materials and methods section.

Besides these two approaches which use diluted samples, a tool which takes fully labeled samples for assessment of the reliability of its result has been published by Bueschl et al. This tool achieves feature reduction, increases the selectivity of compounds with biological origin and assesses molecular structures of measured substances. Nevertheless, the labeling of the samples is time and cost intensive and labeling itself may not be feasible for all samples.

4.2. Improvement/Development

To address bugs, memory as well as speed issues and also accuracy IPO was constantly improved.

4.2.1. Optimization approach

First parameter optimization attempts were based on a screening-modelling-optimization approach (Poster 'Automated XCMS parameter optimization' in the Appendix). The screening step was similar to Zheng et al's approach described in the introduction. The modelling step used a fractional factorial design to again analyze the remaining parameter's influence for significance and to reduce once more the number of parameters to be optimized. Then the optimization step was accomplished using a CCD and evaluated using response surface models. This approach was based on the premise of optimizing peak picking, retention time correction and grouping parameters all together. Therefore the screening and modelling step were reasonable at that time. This first approach used a combination of seven desirability functions used on seven different target values (one of these target values was the number of isotopic peaks). This score was hardly interpretable and due to the desirability function which were only based on individual experience not comprehensible and biased. Developing the scores PPS and RGTV tremendously improved the whole optimization approach. First, due to the nature of the scores splitting the optimization process into optimization for peak picking and optimization for retention time correction and grouping became feasible. This also obviated the need for the screening and modelling step. Second, the new scores were more comprehensible and made the interpretability of the result easier. The first approach also only tested a specific parameter range. Optimal parameters were only estimated within a limited, predefined range. Only after the introduction of the 'maximum focusing' (2.4.5) it was possible to adjust this predefined range and find optimal parameters outside its limits.

4.2.2. IPO's isotope peak identification

With the introduction of PPS the reliability of the isotope peak identification algorithm gained importance. In early versions of IPO the isotope peaks have been identified using the R-package CAMERA. We assumed that isotopes co-elute with their respective parent masses, hence a restriction in the chromatographic axis was possible. This restriction is not implemented by CAMERA which negatively influences the accuracy of the annotation. Also, a manual inspection of the mass deviation of isotopes relative to its ^{12}C mass showed deviation of up to 40 ppm, although a smaller ppm restriction was set. Therefore the mass accuracy was deemed insufficient for our purposes. Furthermore, optimization of big datasets lead to out-of-memory errors and the process of the annotation took quiet long which negatively influenced the time needed for the optimization. To overcome all these issues a new algorithm was implemented which specifically annotates single charged ^{13}C isotopic peaks and moreover is more accurate, faster and does not crash due to lack of memory. The isotope peak identification used the m/z and retention time values to annotate isotopic peaks. This was done by specifying ranges relative to the 'mzmed' and 'rtmed' values of the peak matrix. The size of these ranges could be set by user-specified parameter. In a later version the accuracy of isotope peak identification was improved by adding an intensity check. Therefore the maximum number of carbon atoms of a mass-to-charge ratio was estimated by presuming a hydrocarbon chain. First, the mass-to-charge ratio was divided by the exact mass of a CH3 molecule. This is a molecule which consists of three hydrogen atoms and one carbon atom. Second, the isotope peak intensity was estimated by multiplying the maximum number of carbon atoms with the natural abundance of ^{13}C isotopes and the ^{12}C peak's intensity. Third, the intensity of an isotope peak had to lie within 40% deviation of this intensity. However, this calculation included two errors in reasoning. First CH3 molecules only occur at the ends of a hydrocarbon chain with CH2 molecules in between. Hence, the calculated number of carbon atoms was lower than the theoretical maximum number. Second, spanning a range around the maximum number was not correct either because the calculated ^{13}C isotope peak intensity was the highest possible intensity.

Adding 40% to the upper bound further falsified the calculation. The corrected isotope peak identification algorithm patched these issues assuming the minimum number of carbon atoms being one and additionally had an accurate calculation of the maximum number of carbon atoms.

The first versions of IPO's isotope peak identification algorithm needed parameter settings. Introducing additional parameters to a tool meant for parameter optimization was preposterous. Therefore the algorithm evolved to an identification algorithm without the need of additional parameters. First, by correcting the calculation of the intensity window, a parameter for the intensity range was not needed anymore because the minimum and maximum numbers of carbon already span an intensity range. Second, by using the deviations of the accuracy of the mass spectrometer, which is reflected by the 'mzmin' and 'mzmax' values in the peak matrix, the parameter for the m/z deviation was also rendered redundant. Finally the relative retention time window was fixed with ±0.5% to eliminate the third parameter of the isotope peak identification. Elimination of parameters is on the one hand a good idea on the other hand causes new issues. One issue arises from the m/z window. This window is only accurate when the HRMS is used. The data from mass spectrometer which for example produce only nominal masses does not span an m/z window. Therefore this data cannot be optimized using the IPO isotope peak identification. To overcome this, IPO (version from 1.6) again provides the option to use CAMERA for isotope peak identification. The parameter settings for CAMERA are passed through using the '…'-operator provided by R.

During the development of IPO emphasis was led on fast processing speed. For isotope peak identification a peak has to be checked against all other peaks resulting in a runtime $O(n^2)$. By keeping n low, it's possible to reduce the runtime. Isotope peak identification makes only sense in each file individually; hence separately looking at the peaks of only one file at a time is the logic consequence. Then, IPO orders the sample peaks by their respective m/z values. All peaks with an 'mz' value lower than the sum of the ^{13}C isotope mass and 'mzmax' value of the 250^{th} peaks are checked. This further reduces n and consequently the runtime. R is very fast when calculations can be vectorized. To exploit this, the checks for m/z and retention time are based on

matrices, achieving an additional runtime reduction. Then only for isotopic peaks candidates the intensities are checked using a vectorized approach as well. The found ^{13}C isotope peaks and their respective ^{12}C peak are added to a matrix. Thereafter, the first 250 peaks are removed from the matrix. This process continuous until no peak is left in the peak matrix.

Manual inspection of peaks and their respective isotopic peaks showed that some were very noisy or not fully integrated but denoted as reliable peaks. An example for the latter is presented in **Figure 13**. IPO relies on XCMS to identify real peak-shaped peaks and only uses m/z, retention time and intensity to denote the reliability of these peaks. For additionally assessing the peak shape reliability a simple check was implemented in IPO version 1.6 which is done within the isotope peak detection after the intensity check. This check reads the intensities at 'rtmin' and 'rtmax' from the raw data. A peak is not classified as reliable if either of these two intensities is higher than a third of the maximum peak height. This test is not made by default but can be done using the parameter 'checkBorderIntensity' of the function findIsotopes.IPO(). Similar to the CAMERA parameters it is passed through using the '…'-operator.

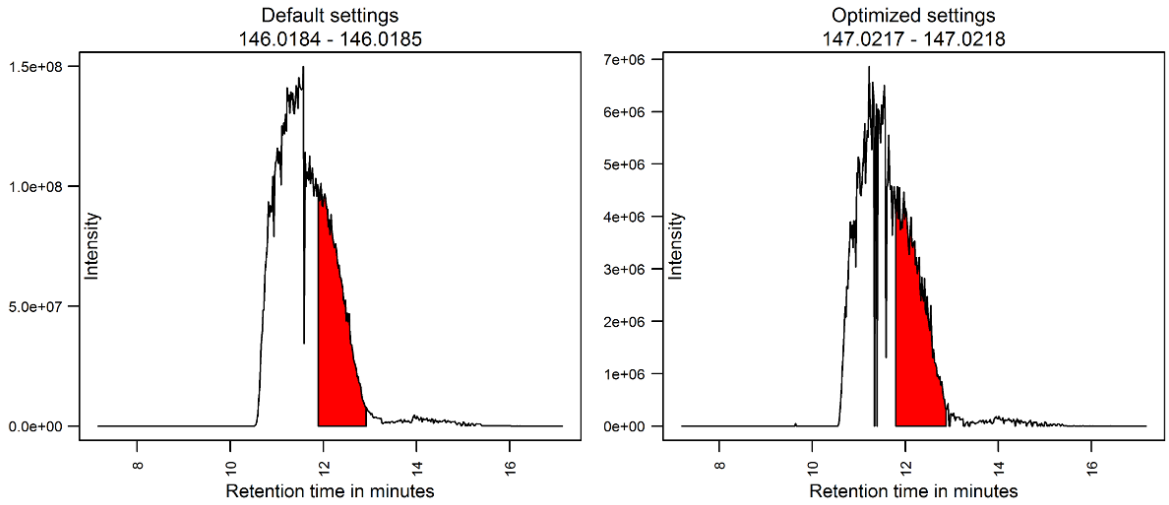


Figure 13: A ^{12}C and its respective ^{13}C peak which would be classified as reliable if the intensity at the start and end points of the peaks is not checked.

^{13}C is the isotope with the highest natural abundance; hence IPO only annotates ^{13}C isotope peaks. However, also other natural, stable isotopes of other chemical elements exist. Annotation of these other isotopes could further improve IPO's reliability but would reduce annotation speed. Furthermore, IPO only identifies single charged isotope peaks. Again, searching for isotope peaks which have multiple charges could increase the reliability but again the annotation time would be increase.

4.2.3. Score development

PPS development was mostly driven by establishing a reliable ^{13}C isotope peak detection algorithm. Early versions of PPS were defined as the ratio of reliable peaks to all other peaks. In the course of the development the influence of the reliable peaks was gradually increased by first using an exponent of 1.5 and then even 2. Also LIPs were defined and the number of all peaks was diminished by these LIPs since they cannot be defined as unreliable peaks.

The RGTV and its respective scores, RCS and GS have undergone some development too. Optimization of retention time correction resulted in errors when using unsuitable values for the 'profStep' parameter. This parameter does always lead to an error using a value bigger than 1. Therefore the IPO-algorithm does only allow a maximum of 1 for this parameter. Also some other settings for 'profStep' seemed to make problems. To generally cope with these problems a penalty was introduced which is given to parameter settings that yield errors.

Additional improvements have been made for the calculation of the grouping score. Early implementations used the ratio of 'reliable' to 'non-reliable groups' without squaring the number of reliable groups. This score had the disadvantage that both 'reliable' and 'non-reliable groups' were reduced as long as the reduction of the 'non-reliable groups' was higher. Therefore, the total number of groups significantly decreased. This contradicted the idea of untargeted metabolomics which aims to detect as much metabolic features as possible to create an accurate metabolomic profile of the samples. To counteract the decrease of groups the number of reliable

groups was squared in the grouping score calculation which increased their recall. Another problem in the calculation of the grouping score was the absence of ‘non-reliable groups’ leading to a ´division by zero´ error. To solve this issue ‘reliable’ and ‘non-reliable groups’ are increased by one if no ‘non-reliable groups’ occur.

Since retention time correction and grouping are optimized simultaneously, the respective scores, RCS and GS, have to be evaluated within the same model. This approach has proven to be problematic because these two scores had to be combined although they often were located in completely different ranges. These different ranges led to different impacts on the response surface model hence favoring the parameters of one of these processing steps. First approaches for solving this problem involved the testing of different weightings. This strategy worked for one experiment but not for others due to the high variation of RCS and GS for different experiments. The solution was to use unity-based normalization. This normalization creates values between zero and one for both scores. After normalization the sum of these two normalized scores is calculated to get RGTV. In contrast to PPS, RGTV cannot be used to determine whether the previous DoE would give a better result. Therefore, the not normalized values of RCS and GS have to be used to detect if a maximum has been found. If either of these two scores stops increasing the optimization of the retention time correction and grouping parameters was finished.

4.2.4. Design of experiment

Until IPO version 1.5.5 a Box-Behnken design was used to create the DoEs. Like the center faced Central-Composite design it tests three levels for each parameter but needs less experiments. The reduction of the number of experiments is achieved by reducing experiments testing the minimum and the maximum levels (36). However, this is not desired when optimizing parameters because the tested parameter levels may lie beyond the minimum or maximum levels tested (see 2.4.5). To sufficiently test these levels the design for DoE creation was changed from BBD to center faced CCD.

4.2.5. Maximum focusing

Before IPO version 1.5.7.1 the ‘contour’ function was used to find the maximum response of given levels. This function gave only the response estimation for two dimensions. For every additional parameter the multidimensional space had to be cut into slices and tested separately. IPO cut each additional dimension into a fixed wide of 0.2 using a recursive function. Using a step wide of 0.2 gave eleven slices for the range -1 to 1. This resulted in $(n-2)^{11}$ contours to be calculated. From IPO version 1.5.7.1 the function ‘expand.grid’ was used instead of the recursive function to increase the performance. The function ‘contour’ uses the function ‘predict’ but with additional calculations which produces an overhead. To improve the performance the ‘contour’ function was replaced directly with ‘predict’.

4.2.6. Parallel computation

Parallelization of separable calculations is a common method to make good use of modern multi-core processors. R provides different packages to do so. IPO (version before 1.5.4.5) used the R-package ‘Rmpi’ for its parallel computation. However, installing the additionally required software is not easily accomplished, especially on Windows. This contradicted IPO’s intention to be useful also for inexperienced users. Therefore parallelization was changed from ‘Rmpi’ to the R-package ‘parallel’. This package is part of R’s core functionality hence does not require additional installation.

4.3. Datasets

Four different datasets which produce different peak shapes and amounts of peaks where used to evaluate the feasibility of the optimization approach itself and the reliability of the parameters suggested by IPO.

4.3.1. HILIC dataset – CARDIONOR

The achieved values for the 'ppm' parameter matched with 30 for nESI and 25 for pESI quiet well the used mass spectrometry resolution of 25,000. The processing of the data with optimized settings showed a striking reduction of the number for peaks for both negative and positive electro spray ionization (Table 11). This reduction was most likely caused by the increased values of the 'peakwidth' parameter. The mean, median and modal peak widths achieved with optimized settings were much higher than those achieved with default settings and better reflected the expected values for metabolomic data from the HILIC method (Table 27). The reduction of the number of peaks can be explained by the use of too small default values for the 'peakwidth' parameter which tend to split a single peak into two or more peaks.

Table 27: Peak width statistic for default and optimized settings for CARDIONOR dataset after peak picking

Peak widths	Default negative ESI	Optimized negative ESI	Default positive ESI	Optimized positive ESI
mean [sec]	39.2	117.6	40.2	109.5
median [sec]	34.9	96.2	34.6	91.2
modal [sec]	38.4	65.1	29.0	67.3

After the grouping very few 'reliable groups' were found for both ionization modes and with optimized as well as default settings (Table 13). Although the optimized settings achieved more reliable peaks, the result was not satisfying. The low number of 'reliable groups' was most likely caused using a dataset which contained all samples for evaluation of the optimization process. Therefore, also blank samples were included which naturally incline to contain only few peaks. For this reason it was to be expected that also with optimized setting only few 'reliable groups' and many 'non-reliable groups' would be achieved. The Venn diagrams in **Figure 14** break the origin of 'non-reliable groups' down to the respective sample classes. The blank sample

class is represented by ‘Bl’, the sample class of the pooled sample injections by ‘QC’ and the serum sample class is labelled by ‘Sa’. The Venn diagram **Figure 14**a shows the origin of ‘non-reliable groups’ for the default settings used on the nESI dataset. It can be seen that all sample classes are present in almost all ‘non-reliable groups’. From all ‘non-reliable groups, 99.7% were found in the blank samples, 99.2% in the ‘Sa’ samples class and 96.9% in the ‘QC’ sample class (Table 28). The Venn diagram **Figure 14**b, which presents the result for nESI with optimized settings, shows that 745 (47.1%) ‘non-reliable groups’ originate only from blank samples. Compared to the default settings the sample class ‘Sa’ produces much less ‘non-reliable groups’ (49.8%). Additionally, it’s reassuring that the sample class ‘QC’ produces very few ‘non-reliable groups’ (15.0%). A similar picture to nESI is given by the Venn diagrams of pESI. Again, all sample classes were present in most ‘non-reliable groups’ (**Figure 14**c). The sample class ‘Bl’ had 3,828 ‘non-reliable groups’ (99.8% of all ‘non-reliable groups’). ‘Non-reliable groups’ originating from the sample class ‘Sa’ were found in 99.2% (3,805) and from ‘QC’ in 96.5% (3,700) of all ‘non-reliable groups’. For the dataset using optimized settings on pESI the ratio of ‘non-reliable groups’ within the sample class ‘Bl’ was 93.5% (Table 28). This sample class alone caused 675 ‘non-reliable groups’ (**Figure 14**d). The sample class ‘QC’ produced a total of 300 (18.6%) ‘non-reliable groups’ and the sample class ‘Sa’ 987 (61.1%).

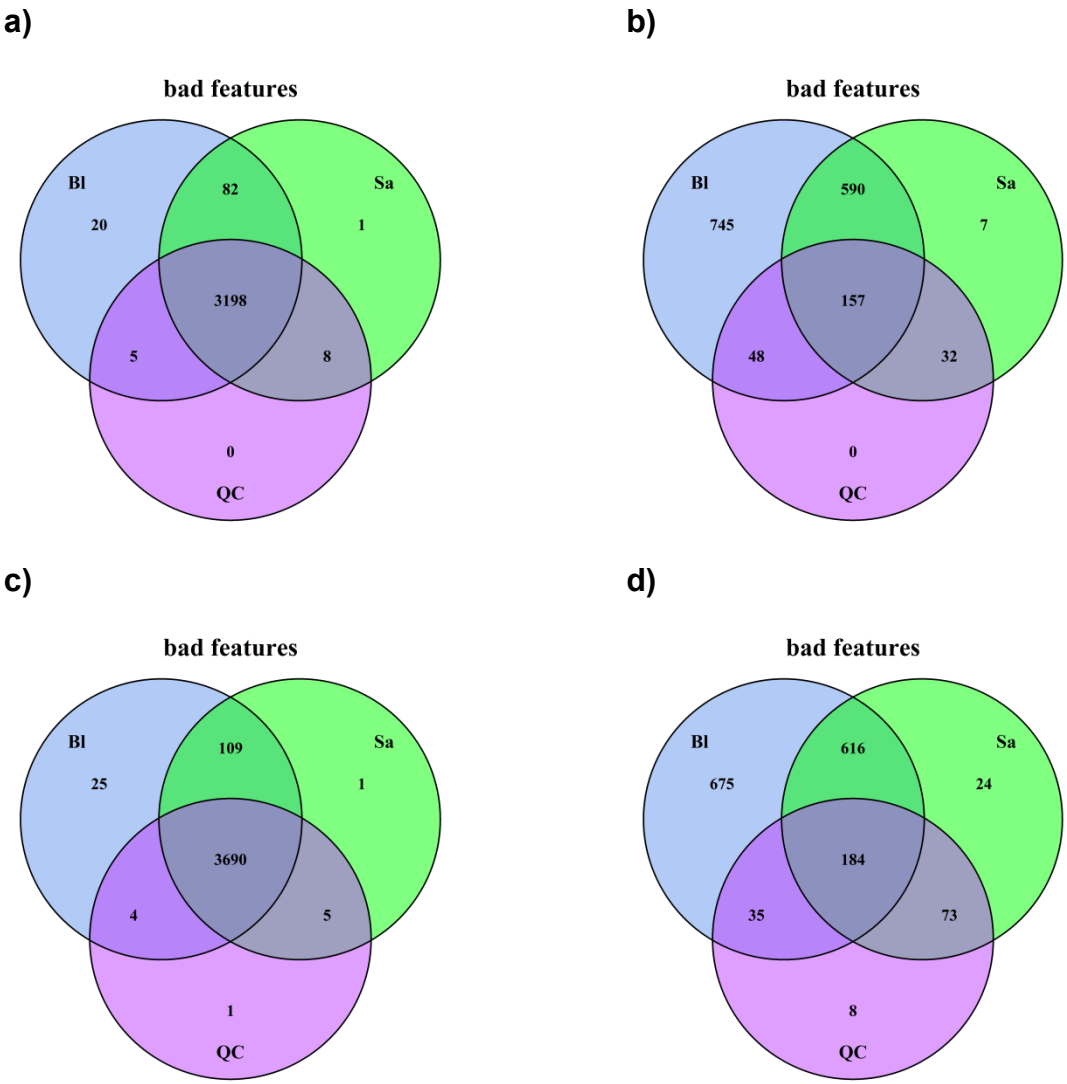


Figure 14: Venn diagrams showing 'non-reliable groups' found within each sample class. The two diagrams on the left steam from default settings and the diagrams on the right from optimized settings. Diagrams a) and b) represent the nESI result; c) and d) show 'non-reliable groups' from pESI.

Table 28: Shows the number of 'non-reliable groups' for each sample class and the percentage compared to all 'non-reliable groups'

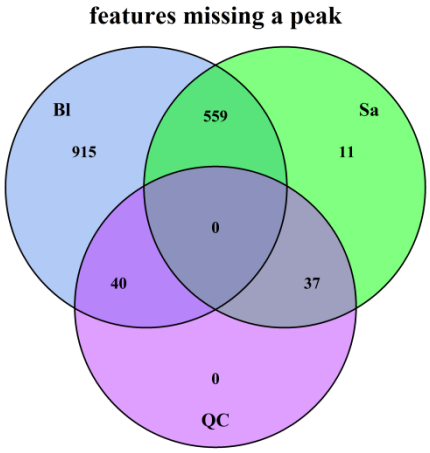
Sample class	nESI		pESI	
	default	opt	default	opt
Bl	3,305 (99.7%)	1,540 (97.5%)	3,828 (99.8%)	1,510 (93.5%)
Sa	3,289 (99.2%)	786 (49.8%)	3,805 (99.2%)	987 (61.1%)
QC	3,211 (96.9%)	237 (15.0%)	3,700 (96.5%)	300 (18.6%)

In the datasets created with optimized setting most 'non-reliable groups' originated from the sample class 'Bl'. But also the sample class 'Sa' yielded many. As argued previously, 'non-reliable groups' can be produced by missing peaks or by samples contributing multiple peaks to the same peak group. Due to the nature of the sample class 'Bl', causing 'non-reliable groups' by missing peaks was to be expected (Table 29). But also the 'Sa' sample class showed missing peaks in 607 and 696 peak groups for nESI and pESI respectively. The 'QC' sample class had the fewest 'non-reliable groups' caused by missing peaks. The missing peaks in the 'Sa' sample class could steam from low concentrations of substances within some samples and therefore no peak could be found in the peak detection step. However, this is the correct conduct and is considered in the processing step after grouping where missing peaks are filled up. Pooling low concentrations into a 'QC' sample further dilutes these concentrations and may be an explanation for the missing peaks in the 'QC' sample class. However, no 'non-reliable groups' due to missing peaks were uniquely caused by the sample class 'QC' as can be seen in **Figure 15**. This also proves the concept of using replicate samples in the optimization approach. Another cause for missing samples could be not perfectly corrected retention time deviations. The optimized 'bw' parameter for nESI and pESI is especially small which supports this theory.

Table 29: Number and percentage of 'non-reliable groups' caused by missing peaks

Sample class	nESI		pESI	
	default	opt	default	opt
Bl	2,999 (90.5%)	1,514 (95.9%)	3,668 (95.6%)	1,497 (92.7%)
Sa	2,120 (64.0%)	607 (38.4%)	2,847 (74.2%)	696 (43.1%)
QC	1,796 (54.2%)	77 (4.9%)	2,342 (61.1%)	117 (7.2%)

a)



b)

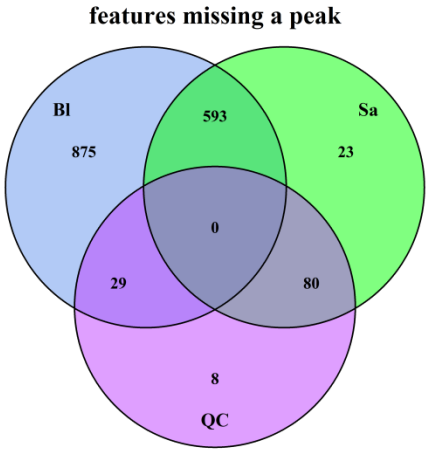


Figure 15: Number of ‘non-reliable groups’ caused by missing peaks in optimized datasets. The dataset nESI is shown in a); dataset pESI in b).

The datasets created with optimized settings showed much less ‘non-reliable groups’ originating from multiple peaks from one sample being grouped than the default datasets (Table 30). The retention time correction does not influence this kind of ‘non-reliable groups’. The most likely reason for the ‘non-reliable groups’ is the grouping parameter ‘bw’ being too large. Additionally, the default peak width settings from the peak picking are very low. This often leads to splitting a peak. In combination with the comparable large, default ‘bw’ parameter level the high number of multiple peaks

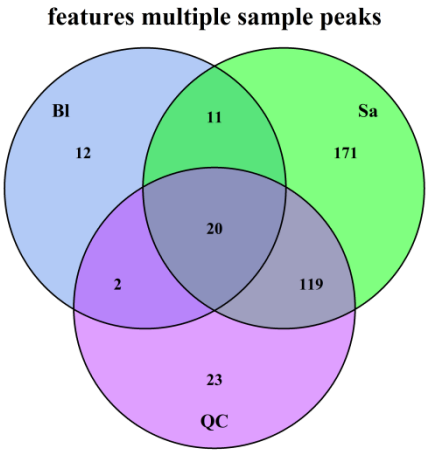
from one sample in the default datasets can be explained. Since the blank samples do not contain many peaks the number of ‘non-reliable groups’ caused by this sample class was relatively low for default settings with 37.3% for nESI and 37.1% for pESI (Table 30). However, the default settings yielded 94.0% and 86.5% ‘non-reliable groups’ in nESI for the sample classes ‘Sa’ and ‘QC’ as well as 93.2% and 84.9% in pESI. These very high numbers could be drastically reduced using the optimized settings. With optimized settings in nESI the ‘non-reliable groups’ were reduced to 2.8%, 20.3% and 10.4% in the sample classes ‘Bl’, ‘Sa’ and ‘QC’ respectively. An equally good reduction was achieved in pESI with ‘non-reliable groups’ being only 1.1% in ‘Bl’, 22.4% in ‘Sa’ and 11.7% in ‘QC’. The most unique ‘non-reliable groups’ came from the sample class ‘Sa’, the fewest from the sample class ‘Bl’ which was to be expected (**Figure 16**).

Although reducing the ‘bw’ parameter could solve this problem it may result in ‘reliable groups’ to turn into bad ones due to missing peaks. Additionally grouping of splitted peaks from the same sample together will have a bad influence on RCS. Therefore, we conclude that RCS is influenced by both retention time correction and grouping parameters, and an appropriate balance between these parameters has to be found which may only be possible with automated optimization.

Table 30: Number and percentage of 'non-reliable groups' caused by multiple peaks from the same sample

Sample class	nESI		pESI	
	default	opt	default	opt
Bl	1,237 (37.3%)	45 (2.8%)	1,421 (37.1%)	17 (1.1%)
Sa	3,116 (94.0%)	321 (20.3%)	3,576 (93.2%)	361 (22.4%)
QC	2,867 (86.5%)	164 (10.4%)	3,256 (84.9%)	189 (11.7%)

a)



b)

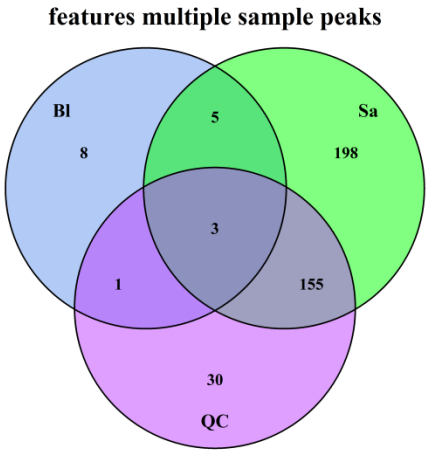


Figure 16: Number of 'non-reliable groups' caused by missing peaks. The optimized dataset nESI is shown in a); the optimized dataset pESI in b).

Finally, it can be concluded that even though the GS did not increase much the optimized settings performed much better than the default setting and the only small increase was caused by including blank samples into the evaluation datasets.

4.3.1.1. Optimization confidence Tests

The summary of the parameter settings achieved in the confidence tests showed outliers for all parameters (Table 14). Most of these outliers were produced by the optimizations which only used two QC injections. This is demonstrated by Table 31 which shows the summaries of the parameters without the optimizations only using two QC injections. It can be seen that no outliers were produced for 1st and 3rd quantile. The only outlier left is the minimum value for the parameter 'bw'. This suggests that using only two out of 36 measurements as basis for parameter optimization decreases the confidence of the optimization result. Although the optimized parameters have a tendency towards expected parameter values the result strongly varies and seems to depend on the chosen measurements. Therefore using a higher number of measurements for the parameter optimization is beneficial and increases the reliability of the achieved parameters.

Table 31: Output of summary for the parameter levels achieved from the optimizations of the confidence test without the 64 optimizations using only two QCs

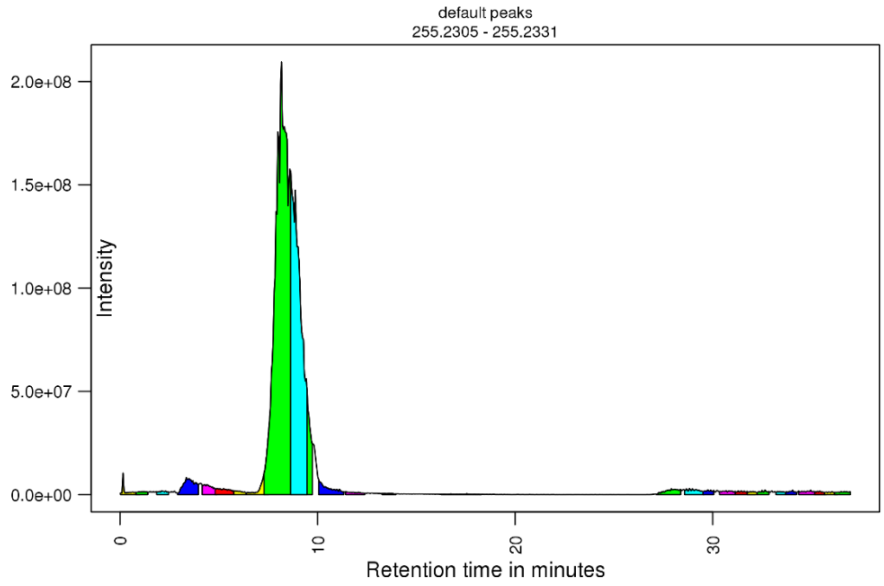
	min PW	max PW	ppm	gapInit	gapExtend	bw	mzwid
Min.	70.0	104.0	18.0	0.000	1.74	0.88	0.003
1st Qu.	86.0	154.0	19.8	0.120	2.10	12.4	0.032
Median	96.4	158.5	21.1	0.220	2.40	12.4	0.039
Mean	92.8	155.0	22.2	0.230	2.27	13.5	0.036
3rd Qu.	100.7	160.0	24.6	0.320	2.40	12.4	0.042
Max.	108.2	187.0	28.5	0.540	2.76	22.0	0.055

4.3.2. HILIC dataset – Bariatric Surgery

The values IPO suggested for the parameters ’min_peakwidth’ and ‘max_peakwidth’ complied well with expected values for the HILIC method. It can be seen that the bariatric surgery dataset showed smaller peak widths than the Cardionor dataset. The bariatric samples were measured about two years after the Cardionor samples. Within this time the HILIC method was modified. One modification was an increased flow rate which could explain the sharper peaks from the bariatric surgery samples. The parameter ‘ppm’ had optimized values of 12.25 for nESI and 11.75 for pESI. These values were to be expected for the resolution of the mass spectrometer which was set to 140,000 (m/z 200) (Table 15). The comparison of the peak picking result showed a striking decrease of the number of peaks (Table 16). This is most likely caused by the increased peak width parameter settings. To small settings for minimum and maximum peak width tend to split a chromatographic peak into multiple peaks. An example for this behavior can be seen in **Figure 17** a) which was generated by default XCMS settings where several peaks are split and hence most peaks are not fully integrated. Compared to **Figure 17** b) which shows peaks

detected by optimized settings, much more peaks are produced by the default settings. It can also be seen that the optimized settings detect the start and end of peaks much better than the default settings.

a)



b)

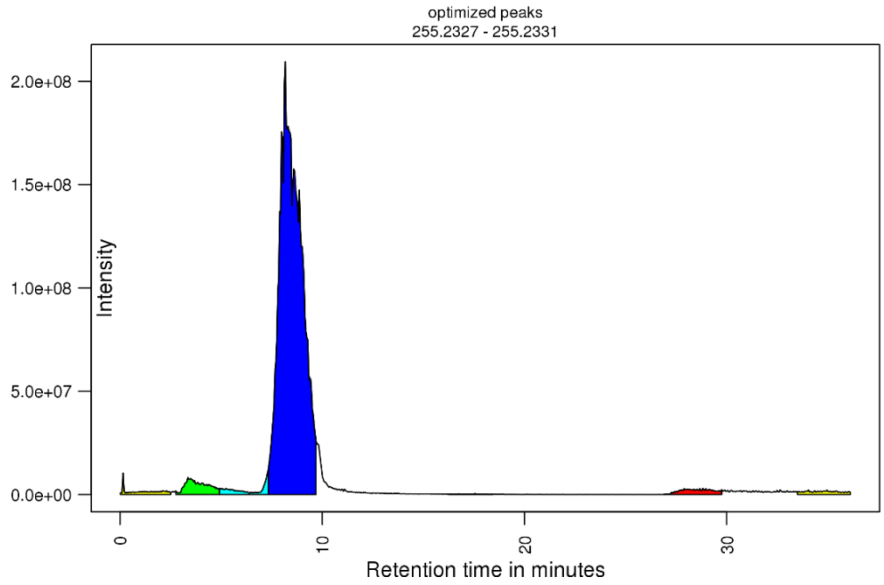
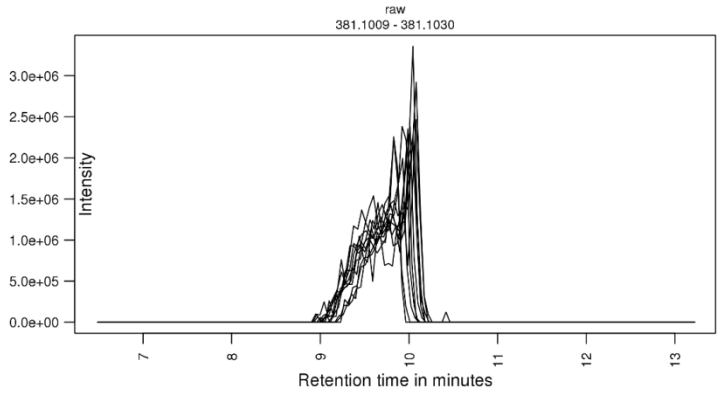


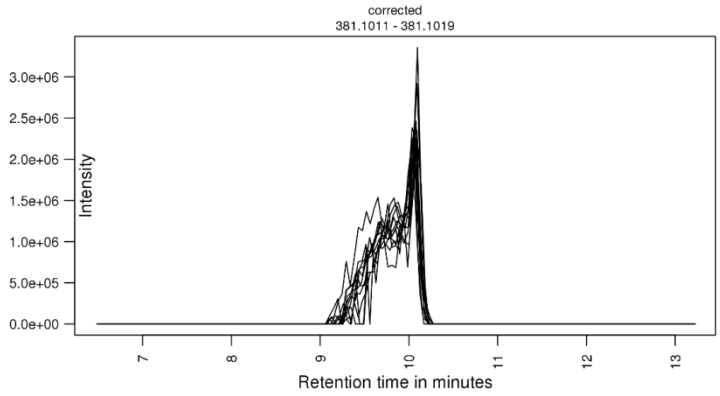
Figure 17: These figures show the same chromatogram. The areas of the peaks detected within the chromatogram are colored. The peaks were detected by a) default settings; b) optimized settings.

The retention time parameters achieved by IPO (Table 17) did not differ much from the default XCMS settings (Table 8). However, RCS significantly increased which was most likely a side effect of reducing the 'non-reliable groups' and increasing the 'reliable groups' (Table 18). However, retention time correction was still necessary which can be seen in **Figure 18**. The chromatogram **Figure 18** a) shows the raw data without retention time correction. It can be seen, that the scans do not align well with each other. The chromatogram **Figure 18** b) shows the data after retention time correction using the default settings and **Figure 18** c) illustrates the same chromatogram after retention time correction with optimized settings. Even by only visually comparing the latter two chromatograms it can be seen that the optimized settings lay near the default parameter settings for retention time correction. Both corrected the retention time shifts very well for the shown m/z range. This suggests that the optimization of the retention time correction might not be responsible for the highly increased RCS. The increased RCS can also be caused by better grouping parameters since peaks from the same sample within a group worsen RCS due to their implicit different retention times. The decrease of 'non-reliable groups' can most likely be explained again by two factors. First, the decrease of the parameter 'bw' compared to the default settings. As discussed earlier a too big setting for this parameter can cause multiple peaks from one sample being grouped together. Second, the increased, optimized parameter 'minfrac' that eliminates groups that don't contain at least one peak from each sample. Also the increased peak picking parameters 'min_peakwidth' and 'max_peakwidth' may have improved the detection of 'reliable groups'. Too small settings for these parameters tend to split a single peak into multiple ones as discussed previously. Compared with a high setting for the grouping parameter 'bw' this can lead to 'non-reliable groups' caused by grouping multiple peaks from the same sample.

a)



b)



c)

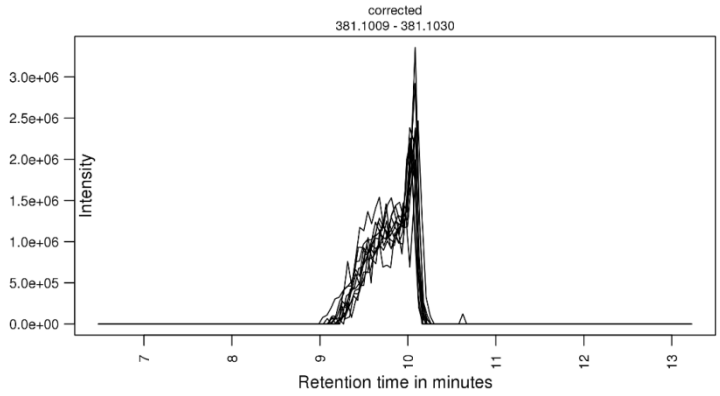


Figure 18: Three chromatograms form the same m/z and retention time range. They show the different effects of retention time correction a) without correction; b) with default settings; c) with optimized settings

4.3.3. RP-HPLC method - Lipidomics

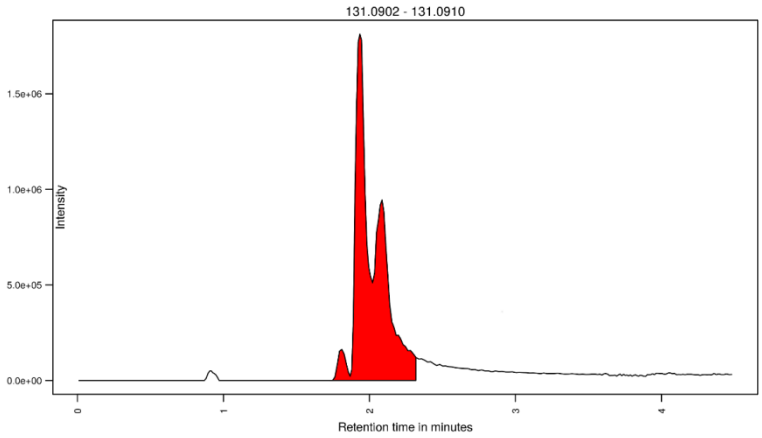
The achieved optimized value for the parameter ‘ppm’ was 7.0. This value was within an expected range for the used resolution of 140.000 of the mass spectrometer. The total number of peaks slightly decreased which may be caused by the increased peak width settings. The optimized retention time correction and grouping values led to a decrease of ‘non-reliable groups’ of 3,213. The value for the parameter ‘bw’ of 0.25 most likely led to the reduction of ‘non-reliable groups’ consisting of multiple peaks from the same sample. The increased number of ‘reliable groups’ on the other hand might be caused by well-chosen retention time correction parameters. The lipidomics method was especially developed to detect a high number of lipids. Lipids can be categorized into fatty acids, glycerolipids, glycerophosphorlipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids and polyketides each category consisting of many metabolites (57). Metabolites within a category often have similar elemental composition but differ in the position as well as number of double bindings as well as the length of carbon chains. These differences result in different retention times which in addition to different mass-to-charge ratios further lead to the detection of individual peaks and makes them distinguishable. Therefore a lot of signals are produced. Although the number of groups identified within the dataset is very high, it may very well reflect the number of lipids as well as possible isotopic peaks within the samples.

4.3.4. IP-RP-HPLC method - Central carbon metabolism

The XCMS peak picking method ‘centWave’ only works well for centroided data. However, the high resolution mass spectrometer Exactive^{TM} Orbitrap system which was used to measure this dataset can only produce profile data. Although raw to mzXML conversion tools like ‘ReadW’ and ‘msConvert’ provide the option to create centroided data out of profile data; this conversation does not always work well. Therefore, the peak picking method ‘matchedFilter’ was chosen for the optimization process. The problem of the ‘matchedFilter’ method is that it cannot adapt well to different peak widths because the parameter ‘fwhm’ only takes one value whereas the

'centWave' algorithm lets define a minimum and maximum peak width. This problem is demonstrated by the two peaks in **Figure 19**. The colored areas denote the peak boundaries found by XCMS. The peak in **Figure 19** a) shows an example where the parameter 'fwhm' was too big for the narrow peaks. Because the 'matchedFilter' algorithm does only use one fixed parameter to define the width of peaks three peaks are falsely classified as a single one. **Figure 19** b) illustrates a peak where the 'fwhm' parameter corresponds well with the actual peak width. Therefore, the peak is integrated within the correct boundaries.

a)



b)

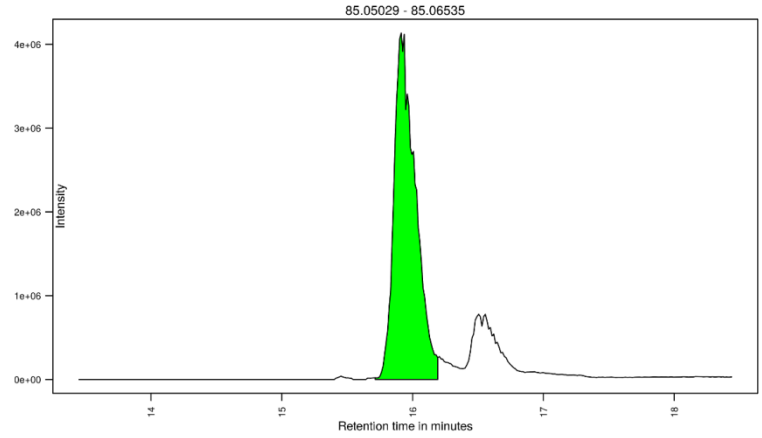


Figure 19: Shows to peaks from the central carbon dataset with a) matching 'fwhm' parameter and peak width; and b) to wide 'fwhm' parameter for the narrow peaks. The colored areas show the peaks integrated by XCMS.

When the ‘matchedFilter’ method was introduced mass spectrometer had a much lower resolution than current state of the art devices. This can also be seen by the default settings this method has. To match resolution of contemporary mass spectrometers the parameters ‘step’ and ‘mzdiff’ were set accordingly. Concluding it can be said, that ‘matchedFilter’ has problems when it comes to LC methods creating different peak widths. However, with the ‘fwhm’ value suggested by IPO almost all peaks were accurately integrated. Also the parameter ‘step’ with a value of 0.0061 reflects well the resolution of the Exactive™ Orbitrap system. The optimization of the retention time correction and grouping parameters performed well too. Almost all ‘non-reliable groups’ could be transformed to ‘reliable groups’. This can only be achieved by reducing existing retention time shifts and choosing proper grouping parameter settings. Although, not many ‘reliable groups’ have been identified a visual inspection of all 190 peak groups showed that no noise was falsely classified as a metabolic feature. Also, from the 1,445 peaks identified 1,140 were grouped to metabolic features. This proves again the reliability of the peak picking settings achieved from the optimization.

5. Bibliography

1. Dunn WB, Ellis DI. Metabolomics: Current analytical platforms and methodologies. TrAC - Trends Anal Chem [Internet]. 2005;24(4):285–94. Available from: http://www.sciencedirect.com/science/article/pii/S0165993605000348

2. Bueschl C, Kluger B, Berthiller F, Lirk G, Winkler S, Krska R, et al. MetExtract: a new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research. Bioinformatics [Internet]. 2012 Mar 1 [cited 2013 Nov 5];28(5):736–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3289915&tool=pmcentrez&rendertype=abstract

3. Bueschl C, Kluger B, Lemmens M, Adam G, Wiesenberger G, Maschietto V, et al. A novel stable isotope labelling assisted workflow for improved untargeted LC–HRMS based metabolomics research. Metabolomics [Internet]. 2014 Dec 4 [cited 2014 Feb 6];10(4):754–69. Available from: http://link.springer.com/10.1007/s11306-013-0611-0

4. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal Chem [Internet]. 2006 Feb 1;78(3):779–87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16448051

5. Benton HP, Wong DM, Trauger SA, Siuzdak G. XCMS 2: Processing Tandem Mass Spectrometry Data for Metabolite Identi cation and Structural Characterization. Anal Chem [Internet]. 2008;80(16):6382–9. Available from: http://pubs.acs.org/doi/abs/10.1021/ac800795f

6. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem [Internet]. 2012 Jun 5;84(11):5035–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3703953&tool=pmcentrez&rendertype=abstract

7. Katajamaa M, Miettinen J, Orešič M. MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics [Internet]. 2006;22(5):634–6. Available from: http://www.biomedcentral.com/1471-2105/6/179

8. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics [Internet]. BioMed Central Ltd; 2010 Jan [cited 2012 Aug 13];11(1):395. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2918584&tool=pmcentrez&rendertype=abstract

9. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak M-Y, et al. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics [Internet]. 2007 Oct [cited 2013 Nov 11];7(19):3470–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17726677

10. Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, et al. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics [Internet]. 2008 Jan [cited 2011 Jun 17];9:163. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2311306&tool=pmcentrez&rendertype=abstract

11. Scheltema R a, Jankevics A, Jansen RC, Swertz M a, Breitling R. PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. Anal Chem [Internet]. 2011 Apr 1;83(7):2786–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21401061

12. Yu T, Park Y, Johnson JM, Jones DP. apLCMS--adaptive processing of high-resolution LC/MS data. Bioinformatics [Internet]. 2009 Aug 1 [cited 2013 Sep 22];25(15):1930–6. Available from: http://bioinformatics.oxfordjournals.org/content/25/15/1930.full

13. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. Bioinformatics. 2006;22(15):1902–9.

14. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M. Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. Curr Bioinform [Internet]. 2012 Mar;7(1):96–108. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3299976&tool=pmcentrez&rendertype=abstract

15. Styczynski MP, Moxley JF, Tong L V., Walther JL, Jensen KL, Stephanopoulos GN. Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. Anal Chem. 2007;79(3):966–73.

16. Creek D j., Jankevics A, Burgess KE V., Breitling R, Barrett MP. IDEOM:anExcelinterfaceforanalysisofLC–MS-basedmetabolomicsdata. Bioinformatics. 2012;28(7):1048–9.

17. Hiller K, Hangebrauk J, Ja C, Spura J, Schreiber K. MetaboliteDetector : Comprehensive Analysis Tool for Targeted and Nontargeted GC / MS Based Metabolome Analysis. Anal Biochem. 2009;81(9):3429–39.

18. Fernández-Albert F, Llorach R, Andrés-Lacueva C, Perera A. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). Bioinformatics. 2014;30(13):1937–9.

19. Eliasson M, Rännar S, Madsen R, Donten MA, Marsden-Edwards E, Moritz T, et al. Strategy for optimizing LC-MS data processing in Metabolomics: A design of experiments approach. Anal Chem [Internet]. 2012 [cited 2013 Aug 1];84(15):6869−6876. Available from: http://pubs.acs.org/doi/abs/10.1021/ac301482k

20. Zheng H, Clausen MR, Dalsgaard TK, Mortensen G, Bertram HC. Time-saving design of experiment protocol for optimization of LC-MS data processing in metabolomic approaches. Anal Chem [Internet]. 2013 Aug 6;85(15):7109–7016. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23841659

21. Uppal K, Soltow Q a, Strobel FH, Pittard WS, Gernert KM, Yu T, et al. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. BMC

Bioinformatics [Internet]. BMC Bioinformatics; 2013 Jan [cited 2013 Feb 5];14(1):15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23323971

22. Brodsky L, Moussaieff A, Shahaf N, Aharoni A, Rogachev I. Evaluation of peak picking quality in LC-MS metabolomics data. Anal Chem [Internet]. 2010 Nov 15;82(22):9177–87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20977194

23. Bajad S, Shulaev V. LC-MS-Based Metabolomics. In: Metz TO, editor. Metabolic Profiling SE - 13 [Internet]. Humana Press; 2011. p. 213–28. Available from: http://dx.doi.org/10.1007/978-1-61737-985-7_13

24. Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol [Internet]. 2004 Nov [cited 2010 Jul 22];22(11):1459–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15529173

25. Deutsch E. mzML: A single, unifying data format for mass spectrometer output. Proteomics. 2008;8:2776–7.

26. Deutsch E. Mass Spectrometer Output File Format mzML. In: Hubbard SJ, Jones AR, editors. Proteome Bioinformatics SE - 22 [Internet]. Humana Press; 2010. p. 319–31. Available from: http://dx.doi.org/10.1007/978-1-60761-444-9_22

27. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. mzML--a community standard for mass spectrometry data. Mol Cell Proteomics. 2011;10(1):R110.000133.

28. mzData [Internet]. Available from: http://psidev.info/index.php?q=node/80#mzdata

29. Danielsson R, Bylund D, Markides KE. Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography-mass spectrometry. Anal Chim Acta. 2002;454(2):167–84.

30. Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics [Internet]. 2008 Jan;9:504. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19040729

31. Arya S, Mount D, Kemp SE, Jefferis G. RANN: Fast Nearest Neighbour Search (wraps Arya and Mount's ANN library) [Internet]. 2014. Available from: http://cran.r-project.org/package=RANN

32. Prince JT, Marcotte EM. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Anal Chem [Internet]. 2006 Sep 1;78(17):6140–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16944896

33. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem [Internet]. 2012 Jan 3;84(1):283–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3658281&tool=pmcentrez&rendertype=abstract

34. Montgomery DC. Design and Analysis of Experiments [Internet]. 8th ed. Wiley; 2012. 478-553 p. Available from: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-EHEP002024.html

35. Box GEP, Behnken DW. Some New Three Level Designs for the Study of Quantitative Variables. Technometrics [Internet]. 1960 Nov [cited 2013 Nov 6];2(4):455–75. Available from: http://www.tandfonline.com/doi/abs/10.1080/00401706.1960.10489912

36. Ferreira SLC, Bruns RE, Ferreira HS, Matos GD, David JM, Brandão GC, et al. Box-Behnken design: An alternative for the optimization of analytical methods. Anal Chim Acta. 2007;597:179–86.

37. Lenth R V. Response-Surface Methods in R , Using rsm. J Stat Softw [Internet]. 2009;32(7):1–17. Available from: http://www.jstatsoft.org/v32/i07

38. Murray KK, Boyd RK, Eberlin MN, Langley GJ, Li L, Naito Y, et al. IUPAC standard definitions of terms relating to mass spectrometry. 2005;85(7):1515–609. Available from: http://eprints.soton.ac.uk/20872/

39. Gold V. International Union of Pure and Applied Chemistry Compendium of Chemical Terminology. 2014;1670. Available from: http://goldbook.iupac.org/PDF/goldbook.pdf

40. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: Open source software for rapid proteomics tools development. Bioinformatics. 2008;24(21):2534–6.

41. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012;30(10):918–20.

42. Medical University of Graz. Multifactorial Treatment of Cardiovascular Risk in Diabetic Patients: Identification of Treatment Non-Responders (CARDIONOR) [Internet]. 2008 [cited 2014 Nov 27]. Available from: http://clinicaltrials.gov/show/NCT00660790

43. Tripolt NJ, Narath SH, Eder M, Pieber TR, Wascher TC, Sourij H. Multiple risk factor intervention reduces carotid atherosclerosis in patients with type 2 diabetes. Cardiovasc Diabetol [Internet]. 2014 [cited 2014 May 27];13(1):95. Available from: http://www.cardiab.com/content/13/1/95

44. Friedrich N. Metabolomics in Diabetes Research. J Endocrinol [Internet]. 2012 Jun 20 [cited 2012 Sep 17];215(1):29–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22718433

45. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. Nat Med [Internet]. 2011 Apr [cited 2013 May 23];17(4):448–53. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3126616&tool=pmcentrez&rendertype=abstract

46. Glauber H, Karnieli E. Preventing type 2 diabetes mellitus: a call for personalized intervention. Perm J [Internet]. 2013 Jan [cited 2014 Feb 11];17(3):74–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3783068&tool=pmcentrez&rendertype=abstract

47. Menni C, Fauman E, Erte I, Perry JRB, Kastenmüller G, Shin S-Y, et al. Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. Diabetes [Internet]. 2013 Dec [cited 2014 Sep 3];62(12):4270–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23884885

48. Daykin C a, Foxall PJD, Connor SC, Lindon JC, Nicholson JK. The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by (1)H nuclear magnetic resonance spectroscopy. Anal Biochem [Internet]. 2002 May [cited 2010 Jul 25];304(2):220–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12009699

49. Bajad SU, Lu W, Kimball EH, Yuan J, Peterson C, Rabinowitz JD. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. J Chromatogr A [Internet]. 2006 Aug 25 [cited 2013 Dec 11];1125(1):76–88. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16759663

50. Arterburn DE, Fisher DP. The Current State of the Evidence for Bariatric Surgery. JAMA [Internet]. 2014;312(9):898–9. Available from: http://jama.jamanetwork.com/article.aspx?articleid=1900490

51. Dixon JB, Blazeby JM. Quality of life after bariatric surgery. lancet Diabetes Endocrinol [Internet]. 2014 Feb [cited 2014 Sep 5];2(2):100–2. Available from: http://www.thelancet.com/journals/a/article/PIIS2213-8587(14)70021-X/fulltext

52. Puzziferri N, Roshek TB, Mayo HG, Gallagher R, Belle SH, Livingston EH. Long-term Follow-up After Bariatric Surgery. Jama [Internet]. 2014 Sep 3 [cited 2014 Sep 3];312(9):934. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2014.10706

53. Yuan M, Breitkopf SB, Yang X, Asara JM. A positive/negative ion-switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. Nat Protoc [Internet]. Nature Publishing Group; 2012 May [cited 2013 Jan 28];7(5):872–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22498707

54. Fauland A, Köfeler H, Trötzmüller M, Knopf A, Hartler J, Eberl A, et al. A comprehensive method for lipid profiling by liquid chromatography-ion cyclotron resonance mass spectrometry. J Lipid Res [Internet]. 2011 Dec [cited 2014 Apr 11];52(12):2314–22. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3220297&tool=pmcentrez&rendertype=abstract

55. Buescher JM, Moco S, Sauer U, Zamboni N. Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. Anal Chem [Internet]. 2010 Jun;82(11):4403–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20433152

56. Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics [Internet]. 2015;16:118. Available from: http://www.biomedcentral.com/1471-2105/16/118

57. Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CRH, Shimizu T, et al. Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res [Internet]. 2009;50 Suppl:S9–14. Available from: http://www.jlr.org/content/50/Supplement/S9.full.pdf+html?sid=9d058055-38de-4bfa-9542-b172f7f4f3b7

Appendix

The Appendix includes the publication of IPO in the journal BMC Bioinformatics, posters the parameter optimization approach presented on various conferences, the vignette as well as the help-files within the IPO R-package.

IPO - Publication in BMC Bioinformatics

Libiseller *et al. BMC Bioinformatics* (2015) 16:118
DOI 10.1186/s12859-015-0562-8

BMC Bioinformatics

SOFTWARE Open Access

IPO: a tool for automated optimization of XCMS parameters

Gunnar Libiseller[1], Michaela Dvorzak[2], Ulrike Kleb[2], Edgar Gander[1], Tobias Eisenberg[3], Frank Madeo[3,4], Steffen Neumann[5], Gert Trausinger[1], Frank Sinner[1,6], Thomas Pieber[1,6] and Christoph Magnes[1*]

Abstract

Background: Untargeted metabolomics generates a huge amount of data. Software packages for automated data processing are crucial to successfully process these data. A variety of such software packages exist, but the outcome of data processing strongly depends on algorithm parameter settings. If they are not carefully chosen, suboptimal parameter settings can easily lead to biased results. Therefore, parameter settings also require optimization. Several parameter optimization approaches have already been proposed, but a software package for parameter optimization which is free of intricate experimental labeling steps, fast and widely applicable is still missing.

Results: We implemented the software package IPO ('Isotopologue Parameter Optimization') which is fast and free of labeling steps, and applicable to data from different kinds of samples and data from different methods of liquid chromatography - high resolution mass spectrometry and data from different instruments.
IPO optimizes XCMS peak picking parameters by using natural, stable ^{13}C isotopic peaks to calculate a peak picking score. Retention time correction is optimized by minimizing relative retention time differences within peak groups. Grouping parameters are optimized by maximizing the number of peak groups that show one peak from each injection of a pooled sample. The different parameter settings are achieved by design of experiments, and the resulting scores are evaluated using response surface models. IPO was tested on three different data sets, each consisting of a training set and test set. IPO resulted in an increase of reliable groups (146% - 361%), a decrease of non-reliable groups (3% - 8%) and a decrease of the retention time deviation to one third.

Conclusions: IPO was successfully applied to data derived from liquid chromatography coupled to high resolution mass spectrometry from three studies with different sample types and different chromatographic methods and devices. We were also able to show the potential of IPO to increase the reliability of metabolomics data.
The source code is implemented in R, tested on Linux and Windows and it is freely available for download at https://github.com/glibiseller/IPO. The training sets and test sets can be downloaded from https://health.joanneum.at/IPO.

Keywords: Metabolomics, XCMS, Parameter optimization, Design of experiments, Isotopologue

Background

Untargeted metabolomics screens biological samples with the aim to reveal new compounds and to understand biological mechanisms. Untargeted metabolomics by using liquid chromatography (LC) generates a huge amount of data when coupled to mass spectrometry (MS). Software packages for automated data processing are needed to successfully process large data sets. Recently, a tool MetExtract has been presented which uses carbon labeling with stable isotopes to find reliable peaks [1,2]. This tool increases the selectivity of compounds with biological origin, performs feature reduction and assesses molecular structures of measured substances. Disadvantages of MetExtract are the time and the cost intensive labeling step and its feasibility which is limited to samples that can be labeled.

A number of software packages for processing LC-MS data have already been developed for data sets of samples that do not rely on labeling [3-12]. They provide methods for peak detection, peak picking, retention time correction and grouping and offer a variety of adjustable parameters to provide reasonable results. But even though these parameters are intended to optimize the

* Correspondence: ca.health@joanneum.at
[1]Joanneum Research Forschungsgesellschaft m.b.H., HEALTH, Institute for Biomedicine and Health Sciences, Graz, Austria
Full list of author information is available at the end of the article

results, wrong parameter selection can lead to distorted outcomes. Parameter optimization is necessary to counter wrong selection. Up to now, several parameter optimization approaches have been proposed to increase the reliability of the results [13-15].

One parameter optimization approach uses design of experiments (DoE) [13]. A designed experiment is a series of tests in which specific modifications are made to the input variables of a process. DoE aims to optimize the response to modifications or to either explain changes of the response variable. . For metabolomics data a dilution series of a pooled sample is measured and a reliability index for each experiment of the DoE is calculated. This reliability index is based on the assumption that peaks which correlate with the dilution series are reliable ones, and those which do not correlate are unreliable peaks. The DoE optimization approach provides quality evaluation of the resulting optimization, but is very time intensive. To accelerate the DoE optimization approach, Zheng, H et al. [14] refined the workflow by first applying a screening step prior to the optimization. Screening steps are usually performed in the first stage of an optimization process with the purpose of identifying the parameters that have large effects on the target variable. For the screening step Zhen et al. used a Plackett-Burman design. Such a fractional factorial design defines only two levels for each parameter and thus requires relatively few experiments. Two levels stand for two different tested values for each parameter. Second, only parameters with a significant positive influence on the target value are optimized and thus the overall optimization time is considerably decreases. However, potential important parameters may be lost because they may fall into a range where they do not significantly influence the target value and hence they may not be further optimized. A software package for parameter optimization which is even faster, widely applicable and free of intricate labeling steps is still missing.

To close this gap we implemented the R-package IPO ('Isotopologue Parameter Optimization') that exploits natural, stable ^{13}C isotope peaks which are ubiquitously present in biological samples. The use of these ^{13}C isotope peaks makes all labeling steps expendable and leads to the calculation of a target value to assess the optimization quality. IPO increases the reliability of peak picking, retention time correction and grouping results and starts the optimization process for the parameters to be optimized at the respective default settings of the XCMS methods and is thus also well suited for inexperienced XCMS users.

Implementation

We developed the R-package IPO to optimize parameters of the open-source package XCMS [3,4]. The process for the parameter optimization by IPO is described in the following subsections (Figure 1).

XCMS parameters

Metabolomics data processing requires peak picking followed by retention time correction and grouping. Multiple methods for each of these steps are provided by XCMS. IPO supports two peak picking, one retention time correction and one grouping method, and can be extended to cover other methods in the future. Various parameters of these methods are optimized by default (Table 1); all other quantitative parameters are optimized only if defined by the user.

The first 'xcmsSet'-method 'centWave' [16] deals with **peak picking**. This is the method of choice for processing centroided data acquired with liquid chromatography (LC) coupled to high resolution mass spectrometry (HRMS). First, 'centWave' identifies regions of interest (ROIs). ROIs are created by combining consecutive centroids within a tolerated m/z deviation, defined by the parameter 'ppm'. Chromatographic peaks are identified within the ROIs using wavelets. The peak width parameters ('min peakwidth' and 'max peakwidth') describe the range of the expected peak widths and determine the scales of the wavelets. The minimum difference of m/z for peaks with overlapping retention times is given by 'mzdiff'.

The second 'xcmsSet'-method 'matchedFilter' [3] also deals with **peak picking**, but it has particularly been developed for low resolution data. In our study, we only optimized high resolution data and therefore we present no example for a parameter optimization with 'matchedFilter'. Nevertheless, IPO also supports this method. The LC-MS data is cut into m/z slices. The widths of these slices are defined by the parameter 'step' and multiple slices can be combined to avoid issues at the boundaries. The parameter 'steps' defines the number of adjacent slices to be combined. Matched filtration is used to filter these slices with a second-derivative Gaussian model peak shape. This Gaussian model peak shape is defined by the parameter 'fwhm'. A signal to noise ratio to filter noisy peaks is determined by the 'snthresh' parameter.

The 'obiwarp' method (Table 1) is responsible for the **retention time correction** [17]. The 'center' parameter indicates the sample which serves as reference sample for retention time correction. If not otherwise specified by the user, XCMS uses the sample with the highest number of peaks as 'center' sample whereas IPO chooses the one with the highest average intensity. First, profiles are generated from the raw data. The parameter 'profStep' defines the widths of these profiles in the m/z dimension. Then, the profiles are compared to each other and a similarity matrix is calculated. Similarity scores are added to recursively generate an optimal path. Off-diagonal transitions are penalized. The parameters 'gapInit' and 'gapExtend' define penalties for gap openings and gap enlargements, respectively.

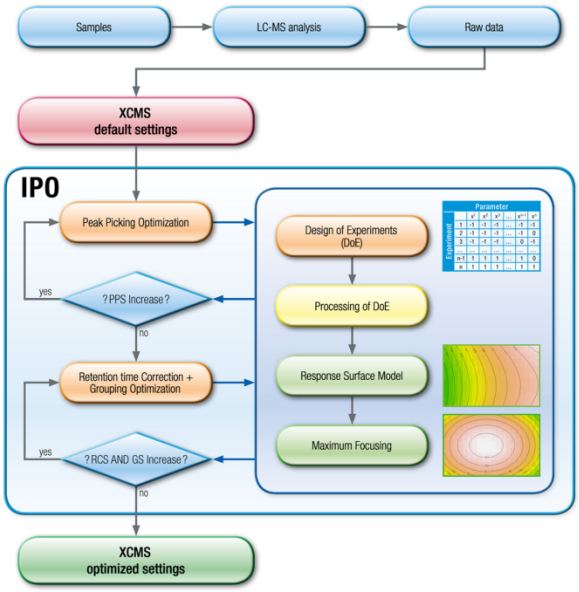


Figure 1 Workflow for the optimization of XCMS parameter. A pooled sample is measured sequentially within the studies. The LC-MS data of the pooled sample are then used for optimization. The DoEs are created by using Box-Behnken designs. The individual experiments of the design are calculated in parallel. Peaks are classified as reliable peaks (RP) when they are part of an isotopologue. These RPs serve as basis for the calculation of the Peak Picking Score (PPS). Two additional scores are introduced for retention time correction and grouping. To improve the quality of retention time correction, the relative retention time deviations within the peak groups are minimized which leads to the Retention time Correction Score (RCS). So called 'reliable groups' and 'non-reliable groups' are defined to assess grouping. The ratio of the squared number of 'reliable groups' to 'non-reliable groups', the Grouping Score (GS), is maximized within the optimization process. The resulting scores are evaluated by using response-surface-models. The combination of parameters that yields the best score is used as new center for the next DoE. The optimization process continues as long as the respective scores are increasing.

The XCMS method 'density' is a method for the **grouping** step. Grouping is the process of combining peaks from different samples with similar masses and retention times to peak groups. The parameter 'bw' is used to define a certain retention time range to find peak groups. 'mzwid' describes the allowed variation in the m/z dimension. The default value for 'mzwid' is 0.25 which is too high for high resolution data and this value was therefore set to 0.025. A valid feature must have a minimum fraction of samples within at least one sample group. This fraction is defined by the parameter 'minfrac'.

Optimization procedure

In general, peak picking is done for each individual data file but for retention time correction and grouping multiple data files are necessary. The optimization procedure splits the parameters by applying a semi sequential approach. Peak picking parameters are optimized first and the retention time correction and grouping parameters are simultaneously optimized afterwards. Grouping results in peak groups by combining peaks with similar masses and retention times from different LC-MS runs. Simultaneous optimization of retention time correction and grouping is necessary because grouping is required

Table 1 XCMS methods and their respective parameters optimized by IPO

XCMS method	Parameters
xcmsSet(method = 'centWave')	min peakwidth, max peakwidth, ppm, mzdiff
xcmsSet(method = 'matchedFilter')	fwhm, step, steps, snthresh, mzdiff
retcor(method = 'obiwarp')	profStep, gapInit, gapExtend
group(method = 'density')	bw, mzwid, minfrac

for the assessment of the retention time correction step, which in turn can improve the grouping. This semi-sequential approach additionally decreases the overall computing time. The different levels for the XCMS parameters are determined by a design of experiments approach [18]. Box-Behnken designs (BBD) serve as basis to generate the DoEs. BBD is a three level incomplete factorial design for fitting a second order response surface model. Three levels denote that for each parameter three different evenly spaced values are tested. The two outer values define a range, the middle value a center point. In contrast to a full factorial design, BBD does not test all factorial combinations, making it highly efficient [19]. For the default levels used by IPO in the first DoE see Additional file 1. To evaluate the result of the DoE, one score for peak picking and one score for retention time correction and grouping is used.

Peak picking

IPO supports the peak picking methods 'centWave' and 'matchedFilter'. By using isotopic peaks it is possible to assess the reliability of peak picking by calculating a peak picking score (*PPS*):

$$PPS = \frac{RP^2}{\text{'all peaks'} - LIP} \quad (1)$$

The PPS is defined as the ratio of reliable peaks (*RP*s) to the number of all peaks (*all peaks*), diminished by the number of 'low intensity peaks' (*LIP*). RP is weighted by the exponential factor 2. Therefore, if the RP value and the number of all peaks increase by the same amount, the PPS increases. This creates an optimization force towards an increased recall of reliable peaks. The exponent value of 2 is an empirical one. The sensitivity for RPs could be enhanced by further increasing this exponent, but then noise would also rise. RPs are defined as peaks that belong to an isotopologue. IPO identifies isotopologues consisting of ^{13}C isotope peaks, which are defined by three criteria. Only peaks that meet all these three criteria are considered isotopic peaks. The tolerable ranges of these criteria are calculated relative to the respective ^{12}C peak. The first criterion states that the mass of the isotope peak has to be within a certain mass range. Second, the isotopic peak must elute at the same time as the parent peak. To restrict peaks on the time axis, a relative retention time window is specified. As a third criterion, the intensities of isotopic peak candidates have to be within a certain intensity window. Therefore, the maximum number of possible carbon atoms (*maxC*) for a specific mass-to-charge ratio presuming a hydrocarbon chain is estimated as follows:

$$maxC = floor\left(\frac{m/z - 2 * CH3}{CH2}\right) + 2 \quad (2)$$

m/z is the mass-to-charge ratio of a peak. *CH2* is the mass of a molecule consisting of one carbon atom and two hydrogen atoms and *CH3* depicts the exact mass of a molecule consisting of one carbon and three hydrogen atoms respectively. First m/z is reduced by 2*CH3 which represent the ends of a hydrocarbon chain. Then, the difference is divided by CH2 which is exemplary for the hydrocarbon bonds within the chain. The function *floor* is used on the result to cut of fractional digits. The previously subtracted 2*CH3 from the ends of the hydrocarbon chain is compensated by + 2 to calculate maxC. Then, intensities of the isotope peaks with one carbon atom and with maxC carbon atoms are estimated by multiplying one and maxC with the natural abundance of ^{13}C isotopes and the ^{12}C peak's intensity. Consequently an intensity window is defined. 'all peaks' includes reliable as well as unreliable peaks. We consider the fact that reliable peaks may exist whose isotope peak concentrations are too low to measure, and would falsely be classified as unreliable ones. To counter this, all peak intensities are arranged in descending order and the average of the lower three percent of the peak intensities is calculated as cut-off value. This cut-off value is used to estimate the sensitivity of the LC-MS system.

For each peak, except for the RPs, the maximum amount of possible carbon atoms is estimated and this amount is then multiplied with the natural ^{13}C isotopic abundance, IA. If the intensity of the peak lies below the cut-off value when multiplied with IA, the peak is neither reliable nor unreliable and is defined as *LIP*.

Retention time correction and grouping

Run-to-run retention time changes have to be corrected. To assess the quality of the retention time correction for one peak group, a group retention time shift (*GRTS*) is calculated as follows:

$$GRTS(x) = \frac{\sum_{n=1}^{k} |(median(x) - x_n)|}{k} \quad (3)$$

x are the retention times of all peaks within one group, k is the number of these retention times and n is an index pointing at the retention time of one individual peak in the peak group. *median(x)* calculates the median value of the retention times for all peaks in one group. For every x the difference to the median retention time is calculated. The average of all these differences is defined as GRTS. The average of all GRTS values yields the average retention time shifts (*ARTS*):

$$ARTS = \frac{1}{k} * sum(GRTS) \quad (4)$$

The number of all GRTSs is defined by k and the function *sum* calculates the sum of these GRTSs. Decreasing the ARTS improves the result. To create a usable optimization value for maximization, the inverse of ARTS is used to define a retention time correction score (*RCS*):

$$RCS = \frac{1}{ARTS} \quad (5)$$

The grouping score (*GS*) is based on the classification of peak groups into 'reliable' and 'non-reliable' ones. '*Reliable groups*' are assumed to show exactly one peak from each injection of a pooled sample. All groups that do not obey this assumption are classified as '*non-reliable groups*'. The absence of a peak within a group can occur due to retention time shifts or due to too low concentrations. GS is calculated as follows:

$$GS = \frac{\text{'reliable groups'}^2}{\text{'non-reliable groups'}} \quad (6)$$

The squared number of 'reliable groups' divided by the number of 'non-reliable groups' is defined as GS. Calculation of the retention time correction and grouping target value (RGTV) is done by the following formula:

$$RGTV = norm(RCS) + norm(GS) \quad (7)$$

To balance the impact of RCS and GS on RGTV, the function '*norm*' is used on RCS as well as on GS. Here, norm is a unity-based normalization used on all RCS values of the experiments of one DoE to scale these values between 0 and 1. The same is done for GS. The normalized values of the same experiments are added giving one RGTV for each experiment of a DoE.

DoE evaluation and adjustment

After the respective scores for each experiment of the DoE have been calculated, response surface models are estimated and applied to evaluate the quality of peak picking, retention time correction and grouping. In a 'maximum focusing step' the combination of parameters which leads to the best respective score is found and used as the new center point for the next DoE. Additionally, in this step, parameter ranges are adjusted according to the following procedure: If the maximum of a parameter shows the same value on the upper and on the lower bound of the parameter range, the range is increased by 20% (zooming out). If the maximum of a parameter has already been located in the middle of the parameter range, with a deviation of less than 25% from the center point, the tool 'zooms in' by narrowing the parameter range by 10% at each bound. The adjusted DoE is recalculated. As long as the respective scores are increasing, this process is continued.

Results and discussion

IPO was applied to untargeted metabolomics data from three different studies that were using different chromatographic devices and methods [20-22]. The sample data originated from human serum, animal tissue (mouse muscle, lung, heart) and yeast samples. All data were high resolution data deriving from LC-HRMS instruments. The three studies used different chromatographic methods that provide data differing in number, shape and quality of the resulting peaks. See Additional file 2 for the characteristics of the data sets. The parameter settings were optimized on training sets and these optimized settings were used on the training and an independent test set. The test set gives an unbiased view of the improvement that can be expected from the approach. The results of the test sets with regard to the parameter optimization steps are presented in Table 2. All response surface models generated during the optimization process of the three data sets are presented in Additional file 3.

Metabolite fingerprinting in human serum (HILIC method)

The metabolite fingerprinting data set used hydrophilic interaction chromatography (HILIC) [20] which typically creates broad peaks. Twelve injections of a pooled sample were used as training set for the parameter optimization and eleven different injections were used as test set. All parameters which were not chosen for optimization were kept at their default values. The PPS of the training set increased by 29% from 1,214 to 1,565 and the PPS of the test set increased by 40% from 1,053 to 1,475. Optimization of the peak picking parameters finished after four DoEs and took about four hours. The number of peaks increased from 55,845 to 57,075 in the training set and decreased from 65,851 to 53,205 in the test set. The number of reliable peaks increased from 6,999 to 8,434 in the training set and from 7,587 to 7,903 in the test set. The optimized peak width parameter lay between 32.2 and 95 seconds. Selected chromatograms, showing the different peak types at distinct masses obtained from the different example data sets are shown in Figure 2. The chromatograms in Figure 2a reveal that the default settings for the peak width parameter can be too small. This results in an only partial integration of the peak, whereas the optimized peak width parameter integrates the peak accurately. The optimization of the retention time correction and grouping parameter finished after five DoEs and 0.8 hours. RCS of the training set increased tenfold by using the optimized settings compared to the RCS of the training set calculated with the default parameters. In the test set the increase of RCS was fifteenfold. The number

Table 2 Results of the example data sets

	Metabolite fingerprinting		Lipidomics		Central carbon metabolism	
pooled sample injections						
training set:	12		4		6	
test set:	11		4		6	
DoEs peakpicking	4		3		2	
DoEs retcor + grouping	5		5		4	
time for peakpicking optimization	3.8 h		1.5 h		0.9 h	
time for retcor + grouping optimization	0.8 h		0.7 h		0.6 h	
overall time	4.6 h		2.2 h		1.5 h	
	default	optimized	default	optimized	default	optimized
#peaks						
training set:	55,845	57,075	33,298	31,710	24,247	24,230
test set:	65,851	53,205	34,415	32,397	27,539	25,609
#RP[a]						
training set:	6,999	8,433	12,606	14,367	2,710	3,351
test set:	7,587	7,903	12,999	14,594	1,582	1,869
#LIP[b]						
training set:	15,497	11,645	15,245	17,284	11,327	11,490
test set:	11,163	10,855	15,643	17,680	12,646	10,962
PPS[c]						
training set:	1,214	1,565	8,802	14,308	568	881
test set:	1,053	1,475	9,001	14,472	168	238
RCS[d]						
training set:	12.3	144.8	67.8	575.4	92.8	311.8
test set:	9.4	142.4	37.6	580.4	48.1	206.7
#reliable groups						
training set:	536	990	3,669	5,343	1,504	2,424
test set:	314	759	1,564	5,639	793	1,855
#non-reliable groups						
training set:	2,636	82	3,605	151	1,217	101
test set:	2,740	70	3,248	110	1,150	69
GS[e]						
training set:	109	11,952	3,734	189,057	1,859	58,176
test set:	36	8,230	753	289,076	547	49,870

[a]reliable peaks; [b]low intensity peaks; [c]peak picking score; [d]retention time correction; score; [e]grouping score

of 'reliable groups' increased from 536 to 990, the number of 'non-reliable groups' decreased from 2,636 to 82 in the training set. In the test set the number of 'reliable groups' increased from 314 to 759 and the number of 'non-reliable groups' decreased from 2,740 to 70.

Lipidomics (RP-HPLC method)

For the lipidomics data set reversed phase high performance liquid chromatography (RP-HPLC) [21] was coupled to a HRMS device. Eight pooled sample injections were analysed. Four of them were used as training set for the optimization process and the four remaining measurements were used as test set. The peak picking parameter 'noise' was set to 20,000. All other parameters were kept at their default values. Optimization of peak picking parameters was finished after three DoEs, which took 1.5 hours. Comparing default to optimized settings the amount of peaks decreased from 33,298 to about 31,710 in the training set and from 34,415 to 32,397 in the test set. The number of RPs increased from 12,606 to 14,367 in the training and from 12,999 to 14,594 in the test set. PPS of the test set increased by 61%, from 9,001 to 14,472. The increase of PPS

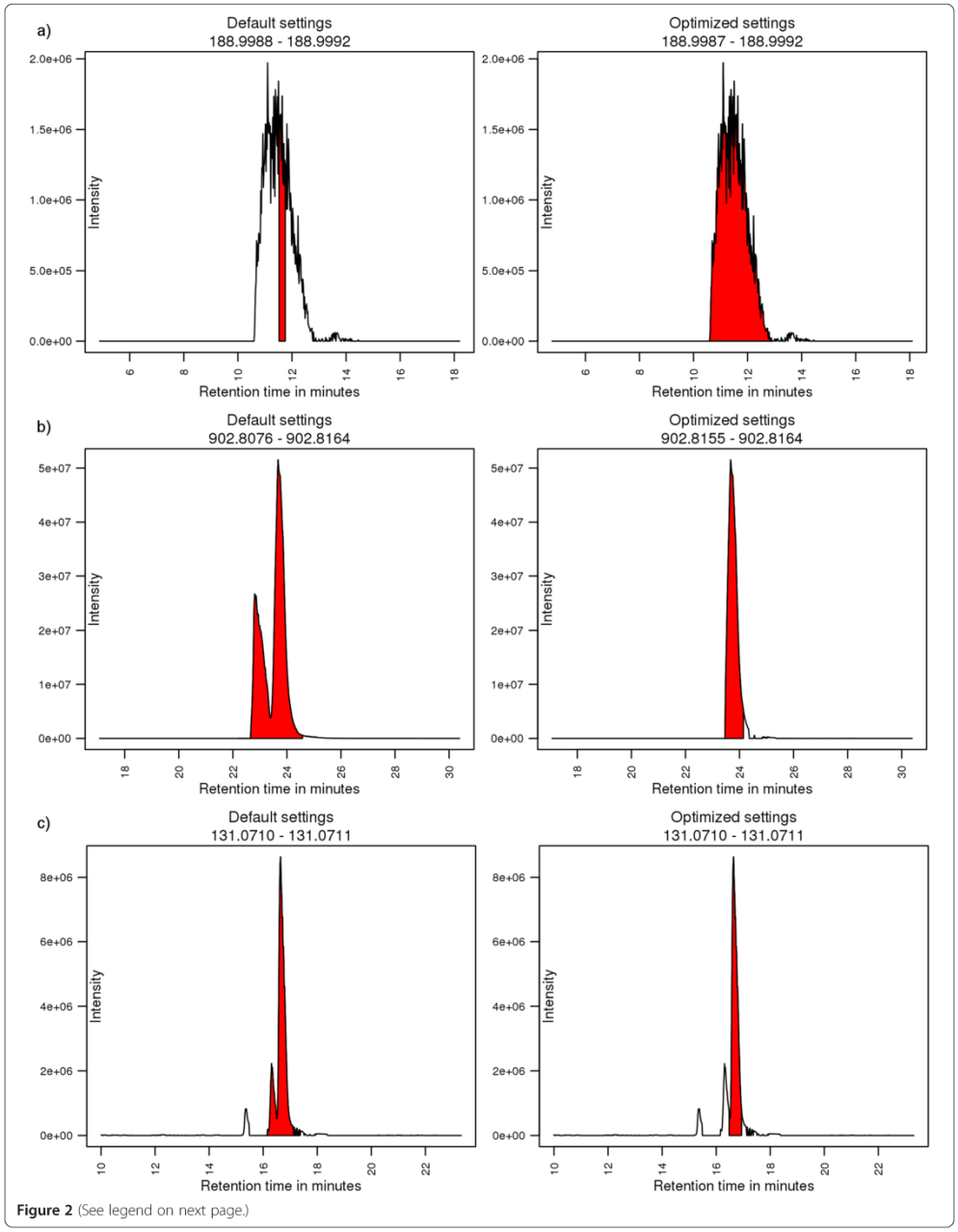


Figure 2 (See legend on next page.)

(See figure on previous page.)
Figure 2 Selected chromatograms showing the different peak types at well-defined masses obtained from the different data sets. Chromatograms derive from **a)** metabolite fingerprinting data set; **b)** lipidomics data set; **c)** central carbon metabolism data set. Peaks derived from default parameters are presented in the left chromatograms and peaks coming from optimized parameters are shown in the chromatograms on the right side, respectively. The peak area integrated by XCMS is colored red. The m/z range for the chromatogram was chosen from the respective minimum and maximum m/z values of the particular peak. Comparison of chromatograms **a)** clearly demonstrate that default peak width parameters were too small for the broad peaks, **b)** shows an example where the mass range used in the default settings was too wide and **c)** illustrate peaks where the default peak width parameters were too wide.

achieved in the training set was 63% from 8,802 to 14,308. The chromatograms in Figure 2b suggest that the default setting for the 'ppm' parameter is too large for peaks generated by HRMS. The optimized parameter results in an m/z range of only 1 ppm for the optimized peak, whereas the default peak spans a range of 9.7 ppm. Parameters for retention time correction and grouping needed 0.7 hours and five DoEs to finish. RCS increased more than eightfold in the training set and fifteenfold in the test set. The amount of 'non-reliable groups' decreased from 3,605 to 151 in the training set and from 3,248 to 110 in the test set. The number of 'reliable groups' increased from 3,669 to 5,343 in the training set and from 1,564 to 5,639 in the test set.

Central carbon metabolism (IP-RP-HPLC method)

The central carbon metabolism data set utilized a modified ion pair-reversed phase-high performance liquid chromatography IP-RP-HPLC [22] method which exhibits an outstanding separation performance, thereby producing very sharp peaks. All parameters that had not been optimized were kept at their default values. Six injections of a pooled sample were used as training set for parameter optimization and six different injections were used as test set. Optimization of peak picking finished after two DoEs and took 0.9 hours. Within the optimization process, the PPS was increased from 568 achieved with the default parameter settings to 881 in the training set and from 168 to 238 in the test set. The chromatograms in Figure 2c show that default settings for the 'peakwidth' parameter are too high for the very sharp peaks generated by this method. The optimization of the retention time correction and grouping parameters for the central carbon metabolism data set finished after four DoEs in 0.6 hours. RCS was more than tripled from 92.8 to 311.8 in the training set and increased fourfold from 48.1 to 206.7 in the test set. 'Non-reliable groups' decreased from 1,217 to only 101 and 'reliable groups' increased from 1,504 to 2,424 which led to a highly increased GS in the training set. In the test set the 'non-reliable groups' decreased from 1,150 to 69 and the 'reliable groups' increased from 793 to 1,855.

The total optimization for the metabolite fingerprinting data set took 3.8 hours, the optimization time for the lipidomics data set took 1.5 hours and the optimization of the central carbon metabolism data set needed 0.9 hours. IPO is also intended to be used by inexperienced users. Therefore, all parameters optimized by IPO start at their respective default values and in a fixed range. Experienced users can further reduce the optimization time by starting with settings closer to their expected parameter values. In general, the results showed that IPO successfully optimized peak picking parameters for data from different LC-methods and different kinds of samples. Peaks coming from the IP-RP-HPLC should be the sharpest of all three studies which is confirmed by the peak width statistics (Table 3). Also, observed peak widths for the metabolite fingerprinting and the lipidomics data sets were in good agreement with the expected peak widths for the respective LC-methods. Especially for broader peaks, the optimized parameters showed a much better peak picking performance than the default settings.

Conclusions

We introduced the software package IPO, 'Isotopologue Parameter Optimization', performing parameter optimization for the open source R-package XCMS. IPO exploits the existence of natural, stable ^{13}C isotopes that are ubiquitous in all biological samples. IPO was applied to LC-HRMS data from tissue, serum and yeast samples and the results showed that it is applicable to data from different types of samples as well as from different LC-MS devices and

Table 3 Peak width parameter settings and resulting peak width statistics of the training sets

	Metabolite fingerprinting		Lipidomics		Central carbon metabolism	
	Default	Optimized	Default	Optimized	Default	Optimized
'peakwidth' parameter [sec]	20-50	32.2-95	20-50	29.6-80	20-50	10-35
mean peak width [sec]	44.2	57.9	44.6	58.4	27.3	15.6
median peak width [sec]	40.6	52.2	41.8	54.5	24.4	12.6
modal peak width [sec]	38.9	51.3	41.4	56.8	10.3	5.8

methods. The optimization time has been remarkable reduced by separating optimization for peak picking parameters from optimization for retention time correction and grouping parameters. IPO is also suitable for XCMS beginners, because the default settings are the start values of the optimization process.

We recommend a powerful workstation with multiple processors and cores, which costs only a fraction of the enormous costs of a modern LC-MS instrument and will enable the user to exploit the full potential of the LC-MS.

IPO is continuously improved, optimization of additional XCMS methods will be implemented, other DoE evaluation techniques will be tested and additional identification of isotopic peaks with the R-package CAMERA [23] will be made available to further increase the power of IPO.

Availability and requirements

Project name: IPO
Project home page: https://github.com/glibiseller/IPO
Operating system(s): Platform independent
Programming language: R
Other requirements: xcms, rsm
License: GNU GPL
Any restrictions to use by non-academics: none

Additional files

Additional file 1: Default levels used in first DoE. The file shows the default levels used by IPO in the first DoE for the different XCMS methods **(Table S1)**.

Additional file 2: Materials. This file contains a detailed description of the three data sets and information on the computation platform used for optimization.

Additional file 3: Response surface models. This file contains the response surface models of all optimization steps of the three data sets.

Abbreviations

ARTS: Average retention time shift; BBD: Box-Behnken design; DoE: Design of experiments; GRTS: Group retention time shift; GS: Grouping score; HILIC: Hydrophilic interaction chromatography; HRMS: High resolution mass spectrometry; IA: Isotopic abundance; IP-RP-HPLC: Ion pair-reversed phase-high performance liquid chromatography; LC: Liquid chromatography; LC-MS: Liquid chromatography coupled to mass spectrometry; LIP: Low intensity peaks; PPS: Peak picking score; RCS: Retention time correction score; RGTV: Retention time correction and grouping target value; ROI: Region of interest; RP: Reliable peaks; RP-HPLC: Reversed phase high performance liquid chromatography.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GL, MD, UK and CM developed the concept of optimization and the optimization scores. GL and SN implemented and tested the R-package. EG, FS and TP were responsible for the acquisition of the data of the metabolomic fingerprint and lipidomics data sets. TE, FM and GT provided data for the central carbon metabolism data set. GL and CM interpreted the optimized results. GL was responsible for formatting and CM and GL for drafting the manuscript which all authors contributed to, critically revised and approved of.

Acknowledgements

The authors thank Beate Boulgaropoulos for critical review and editorial assistance with the manuscript. Additionally we thank Hendrik Treutler and Emanuel Kemmler for testing IPO and providing feedback. This work was supported financially by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), Project Met2Net. FM is grateful to the Austrian Science Fund FWF (Austria) for grants P2349-B12, P24381-B20, I1000 and grant 'SFB Lipotox' and to BMWFW and the Karl-Franzens University for grant 'Unkonventionelle Forschung'. TE is recipient of an APART fellowship of the Austrian Academy of Sciences.

Author details

[1]Joanneum Research Forschungsgesellschaft m.b.H., HEALTH, Institute for Biomedicine and Health Sciences, Graz, Austria. [2]Joanneum Research Forschungsgesellschaft m.b.H., POLICIES, Institute for Economic and Innovation Research, Graz, Austria. [3]Institute of Molecular Biosciences, NAWI Graz, University of Graz, 8010 Graz, Austria. [4]BioTechMed Graz, 8010 Graz, Austria. [5]Department of Stress- and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle, Germany. [6]Department of Internal Medicine, Medical University of Graz, Graz, Austria.

Received: 12 November 2014 Accepted: 30 March 2015
Published online: 16 April 2015

References

1. Bueschl C, Kluger B, Berthiller F, Lirk G, Winkler S, Krska R, et al. MetExtract: a new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research. Bioinformatics. 2012;28:736–8.
2. Bueschl C, Kluger B, Lemmens M, Adam G, Wiesenberger G, Maschietto V, et al. A novel stable isotope labelling assisted workflow for improved untargeted LC–HRMS based metabolomics research. Metabolomics. 2014;10:754–69.
3. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78:779–87.
4. Benton HP, Wong DM, Trauger SA, Siuzdak G. XCMS 2: processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem. 2008;80:6382–9.
5. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem. 2012;84:5035–9.
6. Katajamaa M, Orešič M. Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics. 2005;6:179.
7. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010;11:395.
8. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak M-Y, et al. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics. 2007;7:3470–80.
9. Scheltema R, Jankevics A, Jansen RC, Swertz MA, Breitling R. PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. Anal Chem. 2011;83:2786–93.
10. Lommen A. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem. 2009;81:3079–86.
11. Melamud E, Vastag L, Rabinowitz JD. Metabolomic analysis and visualization engine for LC-MS data. Anal Chem. 2010;82:9818–26.
12. Yu T, Park Y, Johnson JM, Jones DP. apLCMS–adaptive processing of high-resolution LC/MS data. Bioinformatics. 2009;25:1930–6.
13. Eliasson M, Rännar S, Madsen R, Donten MA, Marsden-Edwards E, Moritz T, et al. Strategy for optimizing LC-MS data processing in Metabolomics: A design of experiments approach. Anal Chem. 2012;84:6869–76.
14. Zheng H, Clausen MR, Dalsgaard TK, Mortensen G, Bertram HC. Time-saving design of experiment protocol for optimization of LC-MS data processing in metabolomic approaches. Anal Chem. 2013;85:7109–016.
15. Uppal K, Soltow QA, Strobel FH, Pittard WS, Gernert KM, Yu T, et al. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. BMC Bioinformatics. 2013;14:15.

16. Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics. 2008;9:504.
17. Prince JT, Marcotte EM. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Anal Chem. 2006;78:6140–52.
18. Montgomery DC. Design and Analysis of Experiments. 8th edition. Wiley; 2012. 478-553.
19. Box GEP, Behnken DW. Some New three level designs for the study of quantitative variables. Technometrics. 1960;2:455–75.
20. Bajad SU, Lu W, Kimball EH, Yuan J, Peterson C, Rabinowitz JD. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. J Chromatogr A. 2006;1125:76–88.
21. Fauland A, Köfeler H, Trötzmüller M, Knopf A, Hartler J, Eberl A, et al. A comprehensive method for lipid profiling by liquid chromatography-ion cyclotron resonance mass spectrometry. J Lipid Res. 2011;52:2314–22.
22. Buescher JM, Moco S, Sauer U, Zamboni N. Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. Anal Chem. 2010;82:4403–12.
23. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem. 2012;84:283–9.

Poster – Metabolomics conference, Washington DC, USA, 2012

JOANNEUM RESEARCH HEALTH

A New Integrated Bioinformatics Tool for Metabolomic Data Handling

Libiseller Gunnar [1], Sinner Frank[1], Pieber Thomas[1], Reichelt Wieland[2], Kleb Ulrike[1], Dvorzak Michaela[1], Magnes Christoph[1]

CONTACT

1
JOANNEUM RESEARCH
Forschungsgesellschaft mbH
HEALTH
Institute for
Biomedicine and
Health Sciences

Christoph Magnes

Elisabethstraße 5
8010 Graz, Austria
Phone +43 316 876-4000
Fax +43 316 8769-4000
christoph.magnes@joanneum.at
health@joanneum.at
www.joanneum.at/health

UNI GRAZ

2
University of Graz
Institute of
Molecular Biosciences

Acknowledgements

This work was supported financially by the
Austrian Federal Ministry of Transportation, Innovation and Technology (bmvit), Project Met4CAD
and the
Jubiläumsfond of the Austrian Nationalbank (Proj. 13699 to H.S.).

Literature

[1] ReadW: http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW#Current_Version

[2] XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification; Colin A. Smith, et al; Analytical Chemistry 2006 78 (3), 779–787

[3] Wishart DS, Knox C, Guo AC, et al., HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res. 2009 37 (Database issue): D603–610.

[4] Karp, P.D., Riley, M., Paley, S., and Pellegrini-Toole, A., The MetaCyc Database,Nucleic Acids Research, 30(1):59–61 2002.

[5] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344–D350.

[6] Cotter D, Maer A, Guda C, Saunders B, Subramaniam S. (2006) LMPD: LIPID MAPS proteome database. Nucleic Acids Res. 2006 Jan 1;34 (Database issue): D507–10

[7] Lu, W., Clasquin, M. F., Melamud, E., Amador-Noguez, D., Caudy, A. A., & Rabinowitz, J. D. (2010). Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. Analytical chemistry, 82 (8), 3212–21. doi:10.1021/ac902837x

Introduction

Until now there has been no tool that can handle everything from data conversion to compound identification let alone individualized storage of results, statistical analysis or management of projects, studies and experiments within one graphical user interface – i.e., an integrated bioinformatic tool.

Aim

We designed and construct an integrated bioinformatic tool for

1) **processing of LC/FTMS raw data** generated by metabolomics studies, including peak picking, peak alignment and grouping across study samples by integrating the XCMS package[1].
2) **organizing and archiving the processed raw data** for further statistical analysis and compound identification
3) **identification of compounds in the processed raw data** by accurate mass and retention time using system-specific databases and public compound DBs

Methods

- Programming language: **Java**
- Database: **MySQL**
- Statistical programming language: **R**

Results

We created JOANNEUM RESEARCH Metabolite Database – JRMDb (Fig. 1). The GUI is divided into three parts. In the left one tree shows projects, studies and experiments. The lower part shows a history of actions within a session. In the right part tables and plots can be displayed within tabs. JRMDb features:

- **Semi-automated data conversion** using ReadW[1] (from vendor-file format to standardized mzXML files)
- **Automated peak picking**, peak alignment and grouping of multiple sample batches utilizing the XCMS Package[2]
- **Data management system** to hierarchically organize the processed data from projects, studies and experiments
- **Data filtering tools** (e.g. Blank-, QC-, System-filter) (Fig. 2)
- **Compound identification** by
  - accurate mass and retention time using a system specific data base updated and maintained by the institute
  - accurate mass using public metabolite data bases (e.g. HMDB[3], MetaCyc[4], ChEBI[5], LipidMaps[6])
- **Data visualization tools** (e.g. EIC diagrams (Fig. 3), trend analysis, density plots, feature correlation plots, heatmaps, PCAs).

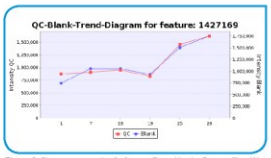


Figure 2: Shows an example of a feature filtered by the System-filter. We periodically mix in pairs of blank- and QC-samples (mix of all samples) during an analysis. If the intensities behave in the same way we assume the feature is just a peak generated by impurities in the system.

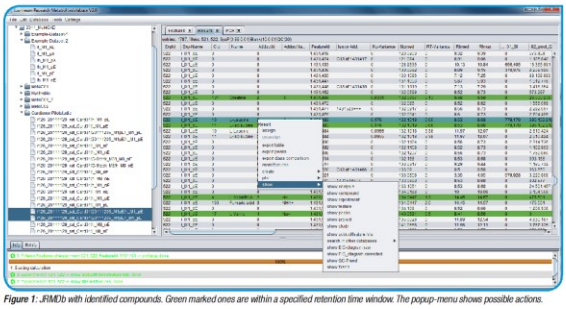


Figure 1: JRMDb with identified compounds. Green marked ones are within a specified retention time window. The popup-menu shows possible actions.

Case Study

In order to challenge the software with a rather scientifically problem we compared typical inducers of autophagy in *Saccharomyces cerevesiae*. The aim was to identify key metabolites which distinguish induction of autophagy by chemical treatment (Rapamycin) and induction through nutrient starvation (HBSS/SN-D/SD-N) (Fig. 4).

The workflow was as following:

- Sample extraction (n = 4 per sample group) using trichloroacetic acid
- QC-sample-preparation by mixing equal amounts of each sample in one vial
- Automated analysis-sequence-generation and sample randomization
- LC/FTMS measurement according to[7]
- Data conversion from vendor file format to mzXML
- Automated XCMS-preprocessing
- XCMS-processing detecting 1203 features
- Storing data to MySQL database
- Data filtration reducing the features to 631
- Compound-identification within the remaining features by m/z and retention time
- PCA-creation (Fig. 5)
- Creation of a heat map (Fig. 6)

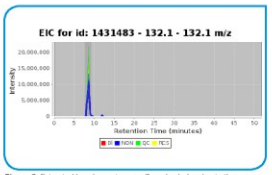


Figure 3: Extracted ion chromatogram. Samples belonging to the same class get automatically the same color.

Conclusion

With JRMDb we have consolidated the metabolomic workflow from data conversion and processing as well as study management to statistical and visual interpretation of the metabolomics data within one program. Due to the modular design further tools like normalization and additional plots can be added easily.

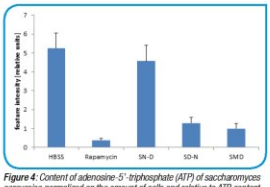


Figure 4: Content of adenosine-5'-triphosphate (ATP) of saccharomyces cerevesiae normalized on the amount of cells and relative to ATP content of cells grown in SMD (synthetic media containing dextrose) in different autophagy inducing growth conditions. HBSS (Hank's balanced salt solution) buffer containing 0,1% glucose; SD-N synthetic media lacking a nitrogen source; SN-D synthetic media lacking dextrose; rapamycin SMD media containing rapamycin.

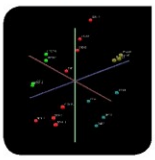


Figure 5: The 3D-PCA clearly clusters the treatment groups from the case study. See also (Fig. 4).

Figure 6: Shows a heat map generated by JRMDb. The intensities of every feature have been normalized to an average of 0 and a standard deviation of 1. Low intensities are represented by a green and high ones by a red color.

THE INNOVATION COMPANY

oef.pbf 12.059

Poster – Metabolomics conference, Glasgow, UK, 2013

JOANNEUM RESEARCH HEALTH

Automated XCMS parameter optimization

A new approach

Libiseller Gunnar[1], Dvorzak Michaela[3], Gander Edgar[1], Kleb Ulrike[3], Narath Sophie H.[1], Pieber Thomas[1,2], Sinner Frank M.[1,2], Sourij Harald[2], Magnes Christoph[1]

JOANNEUM RESEARCH HEALTH

1
JOANNEUM RESEARCH Forschungsgesellschaft mbH
HEALTH
Institute for Biomedicine and Health Sciences

Christoph Magnes

Elisabethstrasse 5
8010 Graz, Austria

Phone +43 316 876-4000
Fax +43 316 8769-4000

health@joanneum.at
www.joanneum.at/health

Medical University of Graz

2
Medical University of Graz
Clinic of Internal Medicine
Division of Endocrinology and Metabolism
Graz, Austria

JOANNEUM RESEARCH POLICIES

3
JOANNEUM RESEARCH Forschungsgesellschaft mbH
POLICIES
Centre for Economic and Innovation Research
Graz, Austria

Introduction

Untargeted metabolomics based on LC/MS relies on automated data processing, such as peak detection, peak picking, retention-time correction or annotation. The software tools available for peak detection and peak picking have many parameters for adjusting algorithms to data set characteristics like MS resolution, chromatographic peak geometry or background noise. The best possible results can rarely be achieved by using the default parameters. Hence, a structured automated workflow to optimize these parameters would be advantageous.

A workflow has been developed that consecutively optimizes all parameters of the subprocedures[1] , and tested in XCMS by using a diluted series of pooled samples[2,3,4].

Aim

Since it is now common to periodically analyze QC samples in LC/MS metabolomic studies, we tried to build a workflow for parameter optimization based on these samples.

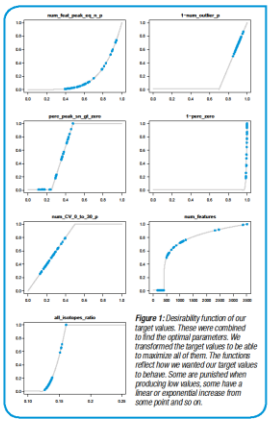


Figure 1: *Desirability function of our target values. These were combined to find the optimal parameters. We transformed the target values to be able to maximize all of them. The functions reflect how we wanted our target values to behave. Some are punished when producing low values, some have a linear or exponential increase from some point and so on.*

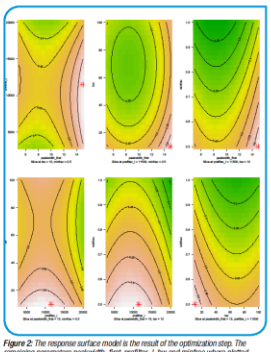


Figure 2: *The response surface model is the result of the optimization step. The remaining parameters peakwidth_first, prefilter_l, bw and minfrac where plotted against each other. The stars indicate the optimal values. For this LC/MS experiment, and the desirability function we choose, the best values would be: peakwidth_first = 15, prefilter_l = 12,000, bw = 10 and minfrac = 0.5. Since some of them are at the outer limits of the parameter settings, these may not reflect the best values, but a good approximation.*

Methods

- Measurements of 6 QC samples analyzed by LC/MS
- Definition of target values which may describe good results (Table 1)
- **Screening**
  - Used to identify which parameters could be important to explaining target values
  - Design of experiment (DoE) done by a Plackett-Burman-Plan (58 experiments)
  - Analysis using a linear model
- **Modeling**
  - Aims to further reduce the number of input parameters
  - DoE by Fractional Factorial design (256 experiments)
  - Analysis using a linear model
- **Optimization**
  - Creation of a desirability function with the remaining target values
  - DoE by Central-Composite-Design (50 experiments)
  - Estimation of response surface model to find optimal settings

Result

The input parameters were reduced from 21 at the beginning of the screening step to 4 at the optimization step (Table 2). We were also able to better evaluate which of our target values may describe reliable peaks.

- **Screening**
  - Reduction of 21 input parameters to 13
  - Only seven target values were influenced by the input parameters and not correlated with other target values
- **Modeling**
  - Reduction of 13 input parameters to 4
  - These 4 input parameters influenced our target values significantly (Table 3)
- **Optimization**
  - Resulting xcms-sets were used to estimate a response-surface-model (Fig. 2)
  - For this LC/MS-experiment, with this desirability function the optimal parameters would be: peakwidth_first = 15, prefilter_l = 12,000, bw = 10 and minfrac = 0.5

Discussion

After the screening step we had to settle for 'centWave' as peak picking method, 'obiwarp' for retention time correction and 'density' for grouping method. This may not have led to optimal results but otherwise the number of necessary experiments would have been much higher. Nevertheless, these experiments helped us to understand the influence of some parameters on our target values.

Outlook

In the future we will conduct experiments to determine whether the number of C13-isotopes relative to the number of peaks would be a good target value for optimization. Fig. 3 shows some trends of this target value for different xcmsSet-parameters-settings. Furthermore, the different steps for xcmsSet generation will be optimized sequentially. This will allow us to test all available methods for peak picking, retention time correction and grouping and make the workflow even faster.

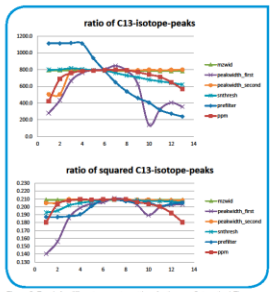


Figure 3: *Trends for different parameter settings for the xcmsSet-method. The upper diagram shows the number of peaks, identified as C13-isotopes, in relation to the number of all peaks; the lower diagram uses the squared number of identified C13-isotopes. We choose these target values, since we think that they might be able to distinguish well between reliable peaks and noise. Some of them also seem to have maxima where an optimization could be beneficial.*

Table 1: *This table shows the target values we had a look at. After the screening step we settled for the 7 most promising.*

Name	Description
num_features	Number of features
num_zeros	Number of peaks with zero intensity
perc_zero	Ratio of peaks with zero intensity relative to num_peaks
num_peaks	Number of peaks
num_peaks_sn_eq_zero	Number of peaks with S/N equal zero, hence from fillPeaks()
num_peaks_sn_gt_zero	Number of peaks with S/N greater zero
perc_peak_sn_gt_zero	Ratio of peaks with S/N greater zero relative to num_peaks
mean_num_peaks_per_feat	Average number of peaks per feature
mean_num_peaks_gt_n	Average number of peaks when less peaks than samples
mean_num_peaks_lt_n	Average number of peaks when more peaks than samples
num_feat_peak_gt_n	Number of features with more peaks than samples
num_feat_peak_lt_n	Number of features with less peaks than samples
num_feat_peak_eq_n	Number of features with as many peaks as samples
std_CV	Standard deviation of coefficients of variation (CV)
mean_CV	Average of CVs
num_CV_0_to_10	Number of CVs ranging from 0 to 10
num_CV_10_to_20	Number of CVs ranging from 10 to 20
num_CV_20_to_30	Number of CVs ranging from 20 to 30
num_CV_30_to_50	Number of CVs ranging from 30 to 50
num_CV_from_50	Number of CVs greater 50
num_outlier	Number of outliers
ratio_C13_isotopes	Ratio of identified C13_isotope_peaks relative to num_peaks
elapsed_runtime	Elapsed runtime of xcmsSet-Generation (just for info)

Table 2: *Tested input parameters. After the screening we decided to use Retcor.obiwarp with no parameters. This reduced the input parameters to 13 and resulted in 256 experiments. For the optimization we only used the four parameters, which influenced our desirability function the most.*

Method	Parameter	Screening	Modeling	Optimization
xcmsSet	ppm	X	X	
	peakwidth_first	X	X	X
	peakwidth_second	X	X	
	snthresh	X	X	
	prefilter_k	X	X	
	prefilter_l	X	X	X
	integrate	X	X	
	mzdiff	X	X	
	profparam_step	X	X	
group	bw	X	X	X
	minfrac	X	X	X
	mzwid	X	X	
	max	X	X	
fillPeaks	fillpeaks	X		
Retcor.obiwarp	profStep	X		
	response	X		
Rector.loess	missing	X		
	extra	X		
	smooth	X		
	span	X		
	family	X		

Table 3: *Influence of the input parameters on our target values. Red background indicates target values to be maximized, blue background indicates minimization. P-values are represented by stars. Blue colored stars, positive algebraic sign: if the input parameter is increased, target value will rise. Red colored stars, negative algebraic sign: inverse behavior of input and output. Last column represents the mean of the other values.*

		num_feat_peak_eq_n_p	num_outlier_p	perc_peak_sn_gt_zero	perc_zero	num_CV_0_to_30_p	num_features	all_isotopes_ratio	D
1	ppm	0	0	***	0	0	***	***	0
2	peakwidth_first	****	****	****	*	****	***	****	****
3	peakwidth_difference	0	0	0	0	0	0	0	0
4	snthresh	**	0	***	0	0	***	0	0
5	prefilter_k	0	0	0	0	0	**	***	*
6	prefilter_l	0	***	0	***	0	****	*	**
7	integrate	0	***	*	0	****	0	***	***
8	mzdiff	0	0	0	0	0	0	0	0
9	profparam_step	0	0	0	**	0	0	*	0
10	bw	****	****	***	***	***	**	0	****
11	minfrac	****	***	****	****	****	****	***	****
12	mzwid	***	0	0	0	0	0	0	0
13	max	0	*	****	***	0	***	0	0
	R²	0.93	0.74	0.93	0.73	0.82	0.78	0.93	0.77

*, **, ***, **** (<0.05, <0.01, <0.001, <2e-16)

References

[1]
Eliasson, M., Rännar, S., Madsen, R., Donten, M.A., Marsden-edwards, E., Moritz, T., Shockcor, J.P., et al. Strategy for Optimizing LC-MS Data Processing in Metabolomics: A Design of Experiments Approach. Anal. Chem, 84(15), 6869–6876 (2012)

[2]
Smith, C.A. and Want, E.J. and O'Maille, G. and Abagyan, R. and Siuzdak, G.: XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification, Analytical Chemistry, 78:779–787 (2006)

[3]
Ralf Tautenhahn, Christoph Boettcher, Steffen Neumann: Highly sensitive feature detection for high resolution LC/MS BMC Bioinformatics, 9:504 (2008)

[4]
H. Paul Benton, Elizabeth J. Want and Timothy M.D. Ebbels Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data Bioinformatics, 26:2488 (2010)

Acknowledgement

This work was supported financially by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), Project Met2Net

THE INNOVATION COMPANY

prmpos13027

Poster – Metabomeeting, London, UK, 2014

Increasing the Reliability of Untargeted Metabolomics by Using Natural Stable ^{13}C Isotopes

Gunnar Libiseller[1], Michaela Dvorzak[2], Ulrike Kleb[2], Alexander Fauland[1], Frank Sinner[1,3], Thomas Pieber[1,3], Christoph Magnes[1*]

CONTACT

1
JOANNEUM RESEARCH Forschungsgesellschaft mbH
HEALTH
Institute for Biomedicine and Health Sciences
***Christoph Magnes**
Neue Stiftingtalstrasse 2
8010 Graz, Austria
Phone: +43 316 876-4201
Fax: +43 316 8769-4201
christoph.magnes@joanneum.at
health@joanneum.at
www.joanneum.at/health

2
JOANNEUM RESEARCH Forschungsgesellschaft mbH
POLICIES
Institute for Economic and Innovation Research
Leonhardstrasse 59
8010 Graz, Austria
policies@joanneum.at
www.joanneum.at/policies

3
Medical University of Graz
Clinic of Internal Medicine
Division of Endocrinology and Metabolism
Graz, Austria

Reference

Fauland A, Köfeler H, Trötzmüller M, et al. (2011) A comprehensive method for lipid profiling by liquid chromatography-ion cyclotron resonance mass spectrometry. J Lipid Res 52:2314–2322. doi: 10.1194/jlr.D016550

Acknowledgement

This work was supported financially by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), Project Met2Net.

Objective

Several tools exist to process the huge amount of data which are produced by untargeted metabolomics. However, the parameter settings for these tools significantly influence the reliability of the result. Here we present a parameter optimization approach to increase the reliability of the result produced by the open source package XCMS.

Methods

50 MCF-7 cell line samples were processed (Fauland et al. 2011) and mixed together to generate a pooled sample. This sample was injected periodically after every third sample into a LC-HRMS device (UHPLC-QExactive) (Fauland et al. 2011). Five of these measurements were then used for the optimization process.

The reliability of the peak picking process was assessed by using natural stable ^{13}C isotopes. Peaks belonging to an isotopologue were classified as **R**eliable **P**eaks (**RP**). If the estimated intensity of a ^{13}C isotope was at the lower limit of quantification, its ^{12}C peak was defined as **L**ow **I**ntensity **P**eaks (**LIP**). The **P**eak **P**icking **S**core (**PPS**) was calculated by:

$$PPS = \frac{\#RP^{1.5}}{\#all\_peaks - \#LIP}$$

The grouping step was needed for the evaluation of the retention time correction result. Therefore retention time correction and grouping were optimized simultaneously. The inverse of the average retention time deviations within the features was defined as a **R**etention time **C**orrection **S**core (**RCS**). This RCS was then used to optimize the retention time correction parameter settings. To optimize grouping parameter settings, features containing exactly one peak from each pooled sample injection were classified as 'good' features, all other features were classified as 'bad' ones. The ratio of the 'good' to 'bad' features resulted in a **G**rouping **S**core (**GS**). Calculation of the Retention time correction and **G**rouping **T**arget **V**alue (**RGTV**) was done by:

$$RGTV = norm(RCS) + norm(GS)$$

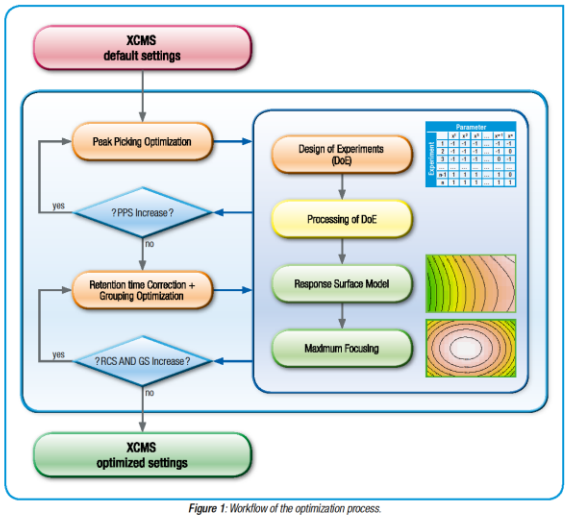


Figure 1: Workflow of the optimization process.

Table 1: Comparison of default and optimized settings

	Default settings	Optimized settings
# peaks	37,197	34,006
#RP[a]	11,263	11,682
#LIP[b]	17,718	17,274
PPS[c]	61.36	75.46
RCS[d]	55.36	482.84
good features	1,687	4,234
bad features	3,056	7
GS[e]	0.55	604.86

[a] Reliable Peaks; [b] Low Intensity Peaks; [c] Peak Picking Score; [d] Retention time Correction Score; [e] Grouping Score

Result

The total number of peaks decreased by 3,191 comparing the default and optimized settings. The number of reliable peaks was increased by 419. This led to a PPS increase of 23 %. The retention time deviations within the features were reduced to less than an eighth. The number of 'good' features increased by 2,547 and only 7 'bad' features appeared using the optimized settings compared to 3,056 achieved with the default settings.

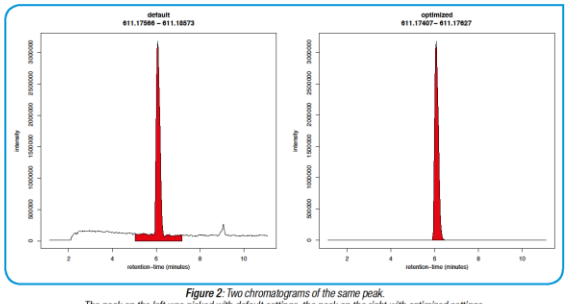


Figure 2: Two chromatograms of the same peak. The peak on the left was picked with default settings, the peak on the right with optimized settings.

THE INNOVATION COMPANY

prm pos 14 052

Poster – Lipidomics conference, Graz, Austria, 2014

A novel approach for automated optimization of untargeted metabolomic data processing

Gunnar Libiseller[1], Michaela Dvorzak[2], Ulrike Kleb[2], Edgar Gander[1], Frank Sinner[1,3], Thomas Pieber[1,3], Christoph Magnes[1*]

CONTACT

[1] JOANNEUM RESEARCH Forschungsgesellschaft mbH
HEALTH
Institute for Biomedicine and Health Sciences
***Christoph Magnes**
Neue Stiftingtalstrasse 2
8010 Graz, Austria
Phone: +43 316 876-4201
Fax: +43 316 8769-4201
christoph.magnes@joanneum.at
www.joanneum.at/health

[2] JOANNEUM RESEARCH Forschungsgesellschaft mbH
POLICIES
Institute for Economic and Innovation Research
Leonhardstrasse 59
8010 Graz
policies@joanneum.at
www.joanneum.at/policies

[3] Medical University of Graz
Clinic of Internal Medicine
Division of Endocrinology and Metabolism
Graz, Austria

Objective

Untargeted metabolomics deals with a huge amount of data. To process this data several tools already exist. However, the parameter settings for these tools significantly influence the quality of the result. To optimize the results we present an approach for automated parameter optimization for the XCMS package.

Methods

Serum samples were mixed together to generate a pooled sample. This sample was injected periodically after every third sample into a LC-HRMS device (HILIC-QExactive). Twelve of these measurements were then used for the optimization process.

To assess the reliability of the peak picking process natural, stable ^{13}C isotopes were used. Peaks belonging to an isotopologue were classified as **R**eliable **P**eaks (**RP**). If the estimated intensity of a ^{13}C isotope was at the lower limit of quantification its ^{12}C peak was defined as **L**ow **I**ntensity **P**eaks (**LIP**). The **P**eak **P**icking **S**core (**PPS**) was calculated by:

$$PPS = \frac{\#RP^{1.5}}{\#all\_peaks - \#LIP}$$

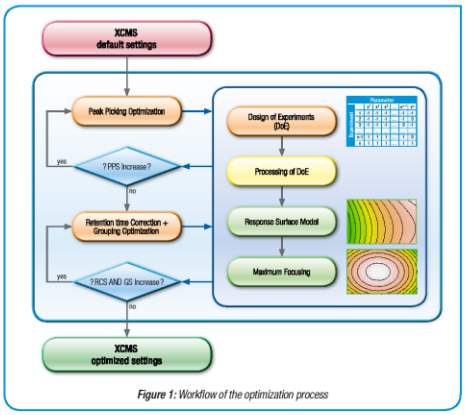


Figure 1: Workflow of the optimization process

Table 1: Comparison of default and optimized settings

	Default settings	Optimized settings
#peaks	66,738	61,911
#RP[a]	8,244	9,288
#LIP[b]	12,227	13,503
PPS[c]	13.73	18.49
RCS[d]	10.75	74.79
good features	361	1,099
bad features	2,725	0
GS[e]	0.13	-

[a]Reliable Peaks; [b]Low Intensity Peaks; [c]Peak Picking Score; [d]Retention time Correction Score; [e]Grouping Score

The grouping step was needed for the evaluation of the retention time correction result. Therefore retention time correction and grouping were optimized simultaneously. The retention time correction was optimized by calculating the inverse of the average retention time deviations within features giving a **R**etention time **C**orrection **S**core (**RCS**). For the grouping step features containing exactly one peak from every pooled sample injection were classified as ‚good', all other features as ‚bad'. The ratio of the ‚good' to ‚bad' features resulted in a **G**rouping **S**core (**GS**). Calculation of the **R**etention time correction and **G**rouping **T**arget **V**alue (**RGTV**) was done by:

$$RGTV = norm(RCS) + norm(GS)$$

Result

The number of peaks decreased by 5,000 with the optimized settings compared to the default settings. The number of reliable peaks increased by over 1,000. PPS increased by 35%. The retention time deviations within the features were reduced to a seventh, the number of 'good' features was tripled and no 'bad' features appeared using optimized settings.

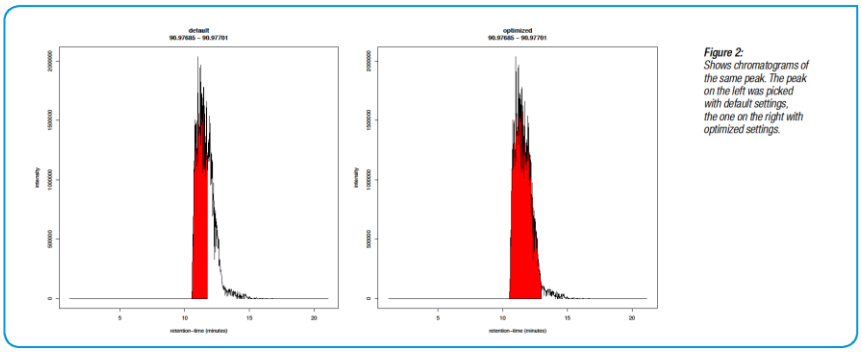


Figure 2: Shows chromatograms of the same peak. The peak on the left was picked with default settings, the one on the right with optimized settings.

Acknowledgement
This work was supported financially by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), Project Met2Net.

THE INNOVATION COMPANY

prmpos14223

Poster – Metabolomics Conference, San Francisco, USA, 2015

JOANNEUM RESEARCH HEALTH

IPO: A Tool for automated XCMS Parameter Optimization

Gunnar Libiseller [1], Michaela Dvorzak [2], Ulrike Kleb [2], Edgar Gander [1], Frank Sinner [1,3], Thomas Pieber [1,3,4], Christoph Magnes [1]

CONTACT

1
JOANNEUM RESEARCH Forschungsgesellschaft mbH
HEALTH
Institute for Biomedicine and Health Sciences
Gunnar Libiseller
Neue Stiftingtalstrasse 2
8010 Graz, Austria
Phone +43 316 876-41 00
Fax +43 316 876-41 00
health@joanneum.at
www.joanneum.at/health

2
JOANNEUM RESEARCH Forschungsgesellschaft mbH
POLICIES
Institute for Economic and Innovation Research
Leonhardstrasse 59
8010 Graz, Austria
Phone +43 316 876-1488
Fax +43 316 876-1480
policies@joanneum.at
www.joanneum.at/policies

3
Medical University of Graz
Department of Internal Medicine
Division of Endocrinology and Metabolism
Auenbruggerplatz 15
8036 Graz, Austria

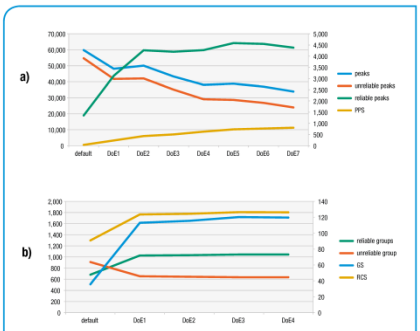
4
CBmed
Center for Biomarker Research in Medicine
Stiftingtalstrasse 5
8010 Graz, Austria

Introduction

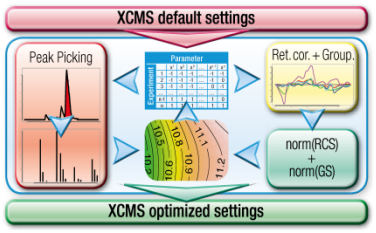
Untargeted metabolomic data processing includes peak detection, retention time alignment and grouping. These tasks are already handled by tools such as XCMS. Different experimental setups require different parameter settings. We aimed to develop a tool for automatic optimization of these parameter settings to increase the reliability of results.

Methods

To ensure compatibility with XCMS the programming language R was chosen. Different parameter settings are obtained by design of experiment. Scores reflecting the reliability of the results are evaluated by using response surface models. The evaluation of the optimization approach was done by using a training and a test set with 15 measurements of a pooled human serum sample for each set.



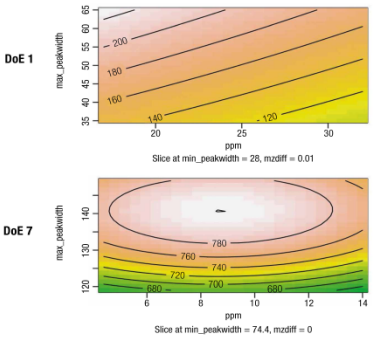
The development of the XCMS-results' characteristics during the optimization process are shown. The diagrams show the development for the optimization of: a) the peak picking parameters; b) the retention time correction and grouping parameters



A simplified scheme of the IPO parameter optimization.

Different parameter settings are chosen by Design of Experiment. For all experiments peak picking is done and a score estimating the reliability of the result is calculated. The scores are then evaluated using a response surface model.

As long as the score increases the best settings are used for the creation of the next DoE. After peak picking the parameters for retention time correction and grouping are simultaneously optimized similarly.



Response surface models of 1st and 7th DoE

Results

IPO, a freely available R-package to optimize XCMS parameters, was developed. By using the parameters obtained from the training set we found an increase of reliable peaks and reliable groups and a decrease of unreliable peaks and groups as well as intra-group retention time deviations in the test set.

Table 1:
The XCMS-methods supported by IPO

Supported XCMS-methods	
peak detection	matchedFilter
	centWave
retention time correction	obiwarp
	loess
grouping	density

Table 2:
Optimized parameters

XCMS-methods	Parameter	Value
peak detection »centwave«	min_peakwidth	66.4
	max_peakwidth	134
	ppm	9.2
	mzdiff	0.0022
retention time correction »obiwarp«	gaplnit	0.448
	gapExtend	2.592
grouping »density«	mzwid	0.01

Table 3:
Comparison of results achieved with default and with optimized settings

	Default settings	Optimized settings	%
peaks	59,326	36,808	-38 %
unreliable peaks	54,143	26,854	-50 %
reliable peaks	1,406	4,552	224 %
PPS	36.5	771.6	2,013 %
RCS	90.0	121.6	35 %
reliable groups	706	1,046	48 %
unreliable group	848	604	-29 %
GS	588	1,811	208 %

THE INNOVATION COMPANY

prm pos 15 307

IPO – Vignette

XCMS Parameter Optimization with IPO

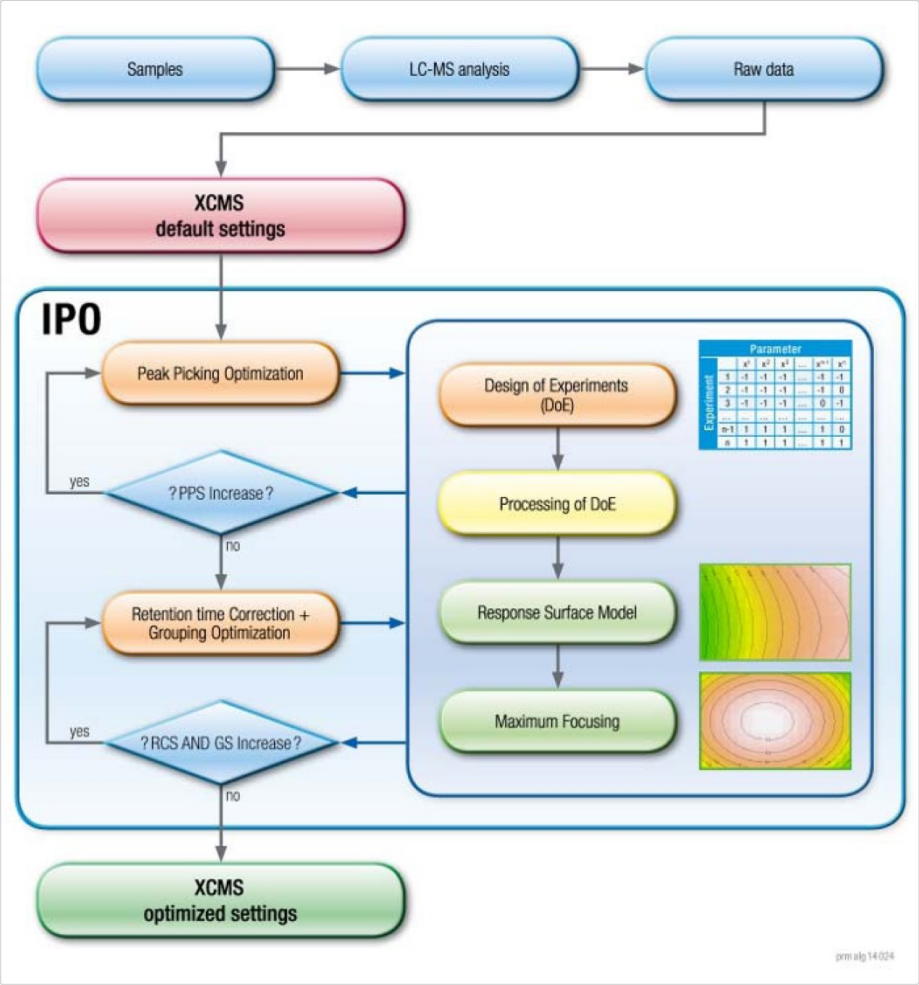
Gunnar Libiseller
Joanneum Research Forschungsgesellschaft m.b.H., Graz, Austria

2015-03-24

Introduction

This document describes how to use the R-package 'IPO' to optimize 'xcms' parameters. Code examples on how to use 'IPO' are provided. Additional to 'IPO' the R-packages 'xcms' and 'rsm' are required. The R-package 'msdata' and'mtbls2' are recommended. The optimization process looks as following:

IPO optimization process



Installation

Install all dependencies for IPO

```
source("http://bioconductor.org/biocLite.R")
biocLite("CAMERA")
biocLite("xcms")
install.packages("rsm")
```

Install RTools

For windows:
Download and install RTools from http://cran.r-project.org/bin/windows/Rtools/
For Unix:
Install the R-development-packages (r-devel or r-base-dev)

Install packages needed for installation from github

```
install.packages("devtools")
```

Install IPO

```
library("devtools")
install_github("glibiseller/IPO")
```

Installing suggested packages

```
#installing suggested packages
biocLite("msdata") #for examples of peak picking parameter optimization
install_github("sneumann/mtbls2") #for examples of optimization of retention time correction and grouping paramet
install.packages("RUnit") #needed for Unittest when checking the package
```

Raw data

'xcms' handles the file processing hence all files can be used that can be processed by 'xcms'.

```
mzdatapath <- system.file("mzData", package = "mtbls2")
mzdatafiles <- list.files(mzdatapath, recursive = TRUE, full.names=TRUE)
```

Optimize peak picking parameters

To optimize parameters different values (levels) have to tested for these parameters. To efficiently test many different levels design of experiment (DoE) is used. Box-Behnken and central composite designs set three evenly spaced levels for each parameter. The method 'getDefaultXcmsSetStartingParams' provides default values for the lower and upper levels defining a range. Since the levels are evenly spaced the middle level or center point is calculated automatically. To edit the starting levels of a parameter set the lower and upper level as desired. If a parameter should not be optimized, set a single default value for 'xcms' processing, do not set this parameter to NULL.

The method 'getDefaultXcmsSetStartingParams' creates a list with default values for the optimization of the peak picking methods 'centWave' or 'matchedFilter'. To choose between these two method set the parameter accordingly.

The method 'optimizeXcmsSet' has the following parameters: - files: the raw data which is the basis for optimization. This does not necessarly need to be the whole dataset, only quality controls should suffice. - params: a list consisting of items named according to 'xcms' peak picking methods parameters. A default list is created by 'getDefaultXcmsSetStartingParams'. - nSlaves: the number ofexperiments of an DoE processed in parallel - subdir: a directory where the response surface models are stored. Can also be NULL if no rsm's should be saved.

The optimization process starts at the specified levels. After the calculation of the DoE is finished the result is evaluated and the levels automatically set accordingly. Then a new DoE is generated and processed. This continues until an optimum is found.

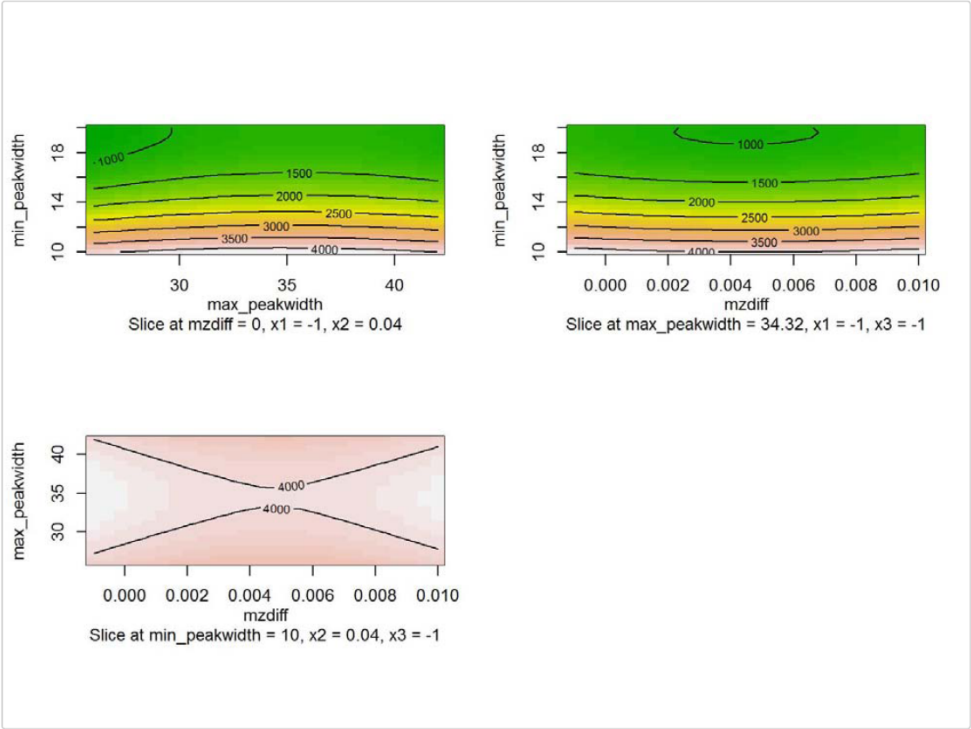
The result of peak picking optimization is a list consisting of all calculated DoEs including the used levels,

design, response, rsm and best setting. Additionally the last list item is a list ('$best_settings') providing the optimized parameters ('$parameters'), an xcmsSet object ('$xset') calculated with these parameters and the response this 'xcms'-object gives.

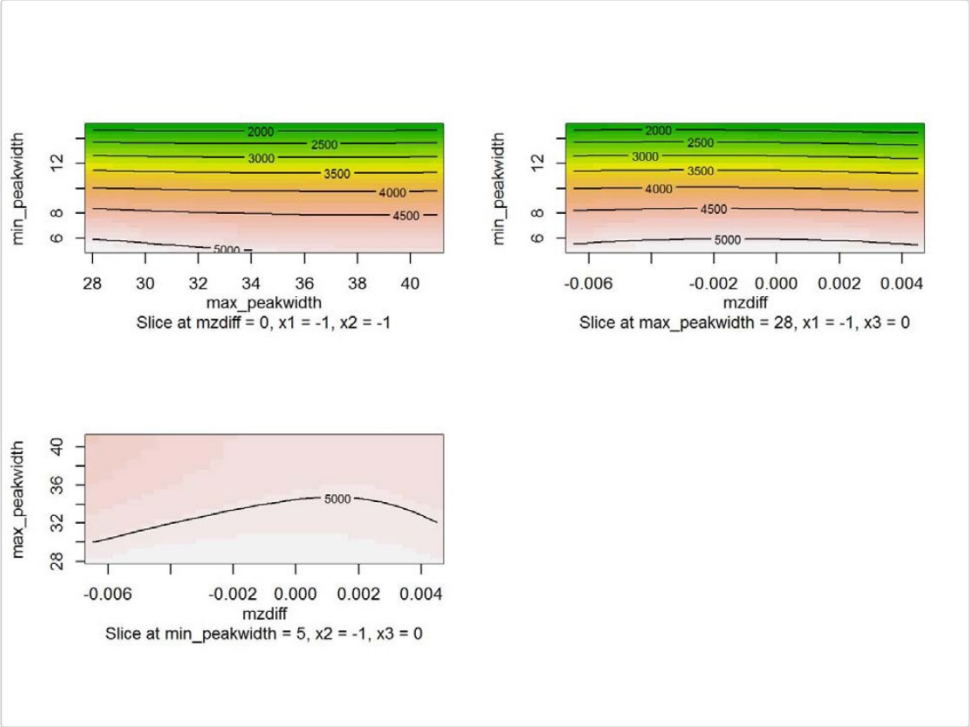
```
library(IPO)

peakpickingParameters <- getDefaultXcmsSetStartingParams('centWave')
#setting levels for min_peakwidth to 10 and 20 (hence 15 is the center point)
peakpickingParameters$min_peakwidth <- c(10,20)
peakpickingParameters$max_peakwidth <- c(26,42)
#setting only one value for ppm therefore this parameter is not optimized
peakpickingParameters$ppm <- 20
resultPeakpicking <- optimizeXcmsSet(files=mzdatafiles[1:4],
                        params=peakpickingParameters, nSlaves=4, subdir='rsmDirectory')
optimizedXcmsSetObject <- resultPeakpicking$best_settings$xset
```

The response surface models of all optimization steps for the parameter optimization of peak picking look as following:

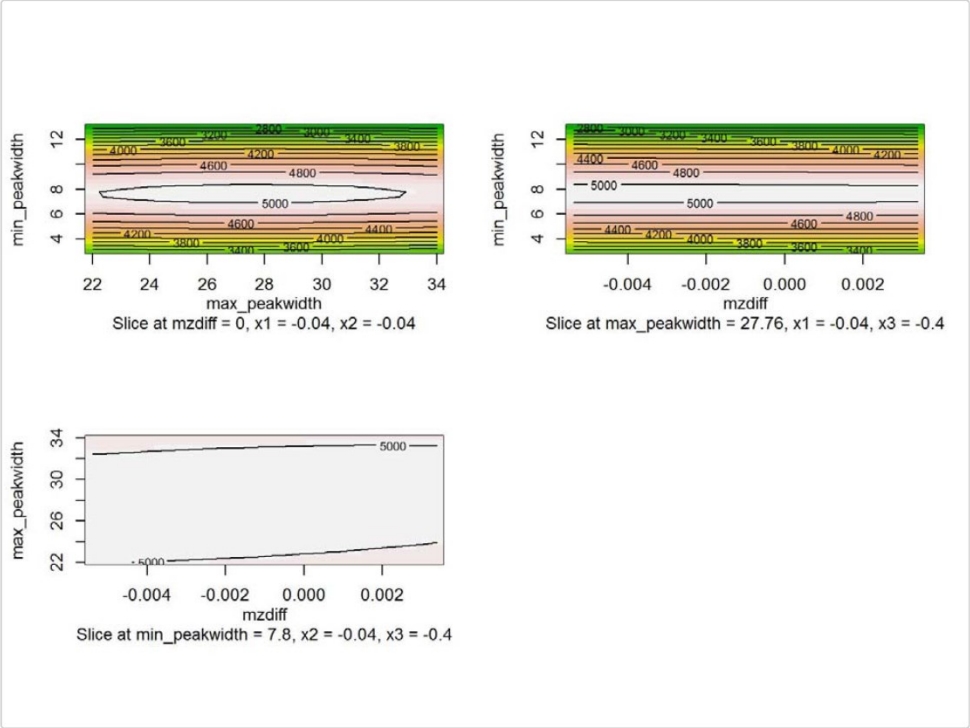
Response surface models of DoE 1 of peak picking parameter optimization



Response surface models of DoE 2 of peak picking parameter optimization



Response surface models of DoE 3 of peak picking parameter optimization



Currently the 'xcms' peak picking methods 'centWave' and 'matchedFilter' are supported. The parameter 'peakwidth' of the peak picking method 'centWave' needs two values defining a minimum and maximum peakwidth. These two values need separate optimization and are therefore split into 'min_peakwidth' and

'max_peakwidth' in 'getDefaultXcmsSetStartingParams'. Also for the 'centWave' parameter prefilter two values have to be set. To optimize these use set 'prefilter' to optimize the first value and 'prefilter_value' to optimize the second value respectively.

Optimize retention time correction and grouping parameters

Optimization of retention time correction and grouping parameters is done simultaneously. The method 'getDefaultRetGroupStartingParams' provides default optimization levels for the 'xcms' retention time correction method 'obiwarp' and the grouping method 'density'. Modifying these levels should be done the same way done for the peak picking parameter optimization.

The method 'getDefaultRetGroupStartingParams' only supports one retention time correction method ('obiwarp') and one grouping method ('density') at the moment.

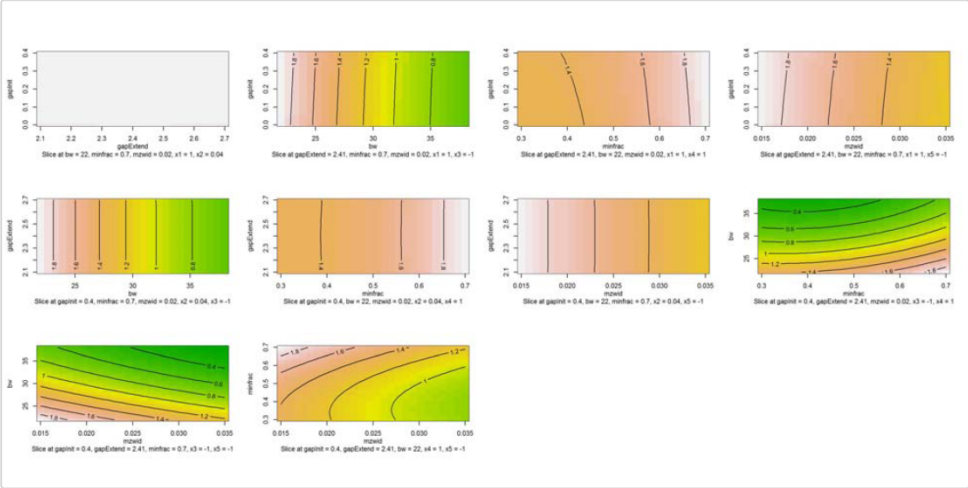
The method 'optimizeRetGroup' provides the following parameter: - xset: an 'xcmsSet'-object used as basis for retention time correction and grouping. - params: a list consisting of items named according to 'xcms' retention time correction and grouping methods parameters. A default list is created by 'getDefaultRetGroupStartingParams'. - nSlaves: the number ofexperiments of an DoE processed in parallel - subdir: a directory where the response surface models are stored. Can also be NULL if no rsm's should be saved.

A list is returned similar to the one returned from peak picking optimization. The last list item consists of the optimized retention time correction and grouping parameters ('$best_settings').

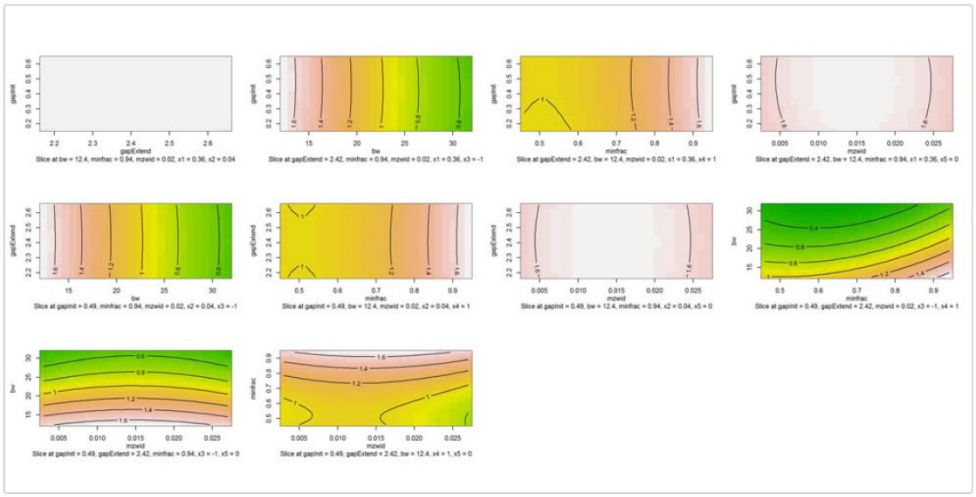
```
retcorGroupParameters <- getDefaultRetGroupStartingParams()
retcorGroupParameters$profStep <- 1
resultRetcorGroup <- optimizeRetGroup(xset=optimizedXcmsSetObject, params=retcorGroupParameters,
                      nSlaves=4, subdir="rsmDirectory")
```

The response surface models of all optimization steps for the retention time correction and grouping parameters look as following:

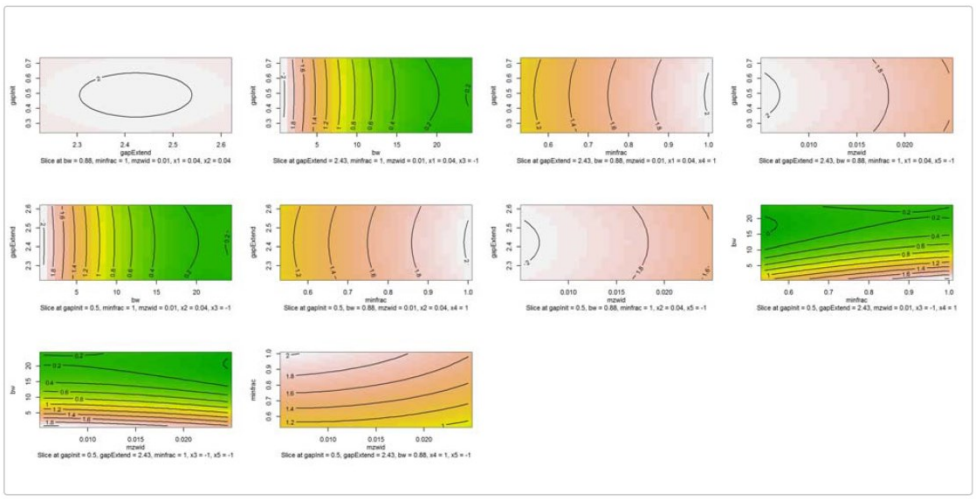
Response surface models of DoE 1 of retention time correction and grouping parameter optimization



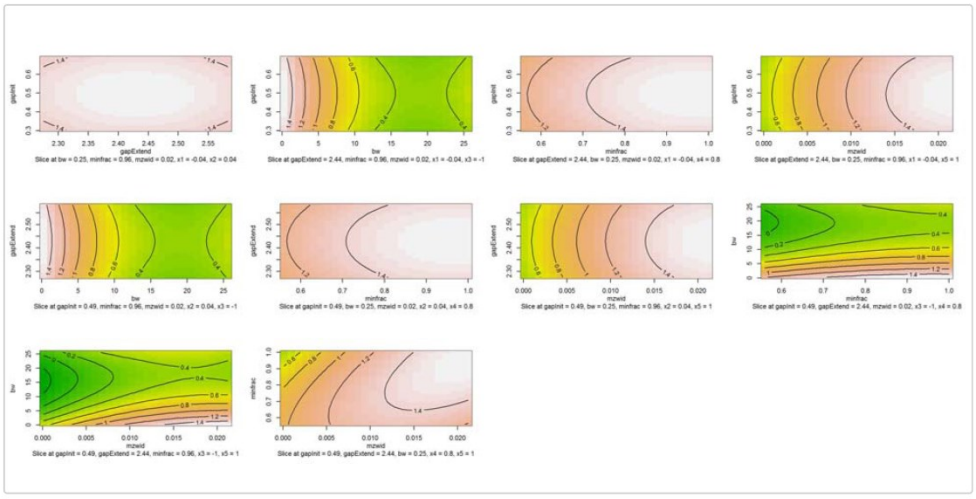
Response surface models of DoE 2 of retention time correction and grouping parameter optimizationn



Response surface models of DoE 3 of retention time correction and grouping parameter optimization



Response surface models of DoE 4 of retention time correction and grouping parameter optimization



Currently the 'xcms' retention time correction method 'obiwarp' and grouping method 'density' are supported.

Display optimized settings

A script which you can use to process your raw data can be generated by using the function 'writeRScript'.

```
writeRScript(resultPeakpicking$best_settings$parameters, resultRetcorGroup$best_settings, nSlaves=4)
```

IPO – help files

Package 'IPO'

May 11, 2015

Type Package

Title Automated Optimization of XCMS Data Processing parameters

Version 1.7.0

Date 2015-05-07

Author Libiseller, Gunnar <Gunnar.Libiseller@joanneum.at>
Magnes, Christoph <christoph.magnes@joanneum.at>

Maintainer Libiseller, Gunnar <Gunnar.Libiseller@joanneum.at>

Depends xcms, rsm, CAMERA

Suggests RUnit, BiocGenerics, msdata, mtbls2

Description The outcome of XCMS data processing strongly depends on the parameter settings. IPO (`Isotopologue Parameter Optimization`) is a parameter optimization tool that is applicable for different kinds of samples and liquid chromatography coupled to high resolution mass spectrometry devices, fast and free of labeling steps. IPO uses natural, stable 13C isotopes to calculate a peak picking score. Retention time correction is optimized by minimizing the relative retention time differences within features and grouping parameters are optimized by maximizing the number of features showing exactly one peak from each injection of a pooled sample. The different parameter settings are achieved by design of experiment. The resulting scores are evaluated using response surface models.

License GPL (>= 2) + file LICENSE

URL https://github.com/glibiseller/IPO

biocViews Metabolomics, MassSpectrometry

R topics documented:

`IPO-package` *Automated Optimization of Untargeted Metabolomics LC-MS Data Processing*

Description

IPO provides a framework for parameter optimization for the software package XCMS. It provides optimisation of peak picking parameters by using natural, stable 13C isotopes. Retention time correction is optimized by minimizing the relative retention time differences within features and grouping parameters are optimized by maximizing the number of features showing exactly one peak from each injection of a pooled sample.

Details

An overview of how to use the package, including the most important functions

Author(s)

Gunnar Libiseller

Maintainer: gunnar.libiseller@joanneum.at

References

Lenth, R. V. (2009). Response-Surface Methods in R , Using rsm. Journal of Statistical Software, 32(7), 1-17. Retrieved from http://www.jstatsoft.org/v32/i07

Smith, C.A. and Want, E.J. and O'Maille, G. and Abagyan,R. and Siuzdak, G.: XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification, Analytical Chemistry, 78:779-787 (2006)

Ralf Tautenhahn, Christoph Boettcher, Steffen Neumann: Highly sensitive feature detection for high resolution LC/MS BMC Bioinformatics, 9:504 (2008)

H. Paul Benton, Elizabeth J. Want and Timothy M. D. Ebbels Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data Bioinformatics, 26:2488 (2010)

Yu, H. (2002). Rmpi: Parallel Statistical Computing in R. R News, 2(2), 10-14. Retrieved from http://cran.r-project.org/doc/Rnews/Rnews_2002-2.pdf

See Also

xcms

Examples

```
library(msdata)
library(IPO)

mzmlfile <- file.path(find.package("msdata"), "microtofq/MM14.mzML")

paramsPP <- getDefaultXcmsSetStartingParams()
paramsPP$mzdiff <- -0.001
paramsPP$min_peakwidth <- c(7,14)
paramsPP$max_peakwidth <- c(20,30)
resultPP <- optimizeXcmsSet(mzmlfile, paramsPP, subdir="microtofq")
#resultRG <- optimizeRetGroup(resultPP$best_settings$xset)

#writeRScript(resultPP$best_settings$parameters, resultRG$best_settings, 4)
```

---

attachList *Attaching one list at the end of another*

Description

This function attaches one list at the end of another list.

Usage

```
attachList(params_1, params_2)
```

Arguments

params_1	A List
params_2	A second list which will be attached at the end of the first list.

Details

This is a convenience funktion, but the implementation is not optimized for speed.

Value

A List composed of the two input lists.

Author(s)

Gunnar Libiseller

Examples

```
a <- list("a"=1, "b"=2)
b <- list("c"=4, "d"=4)
attachList(a, b)
```

`calcPPS` *Calculation of a peak picking score (PPS) by using natural, stable 13C isotopes*

Description

This function calculates PPS by identifying natural, stable 13C isotopes within an xcmsSet object. Peaks beeing part of an isotopologue are defined as reliable peaks (RP). Peaks which are not part of an isotopologue and where the intensity of possible isotopes is below a cutoff are defined as 'low intensity peaks' (LIP). PPS is then calculated by:
PPS = RP^2/(#all_peaks - LIP)

Usage

```
calcPPS(xset, isotopeIdentification, ...)
```

Arguments

`xset`	`xcmsSet` object
`isotopeIdentification`	This parameter defines the method for isotope identification. The method IPO was especially implemented for high resolution data. CAMERA is an established isotope and adduct annotation package.
`...`	Additional parameters to CAMERA's findIsotopes function.

Details

Calculation of a peak picking score (PPS) by using natural, stable 13C isotopes

Value

An array with 5 items:
1. Space for experimentid of the Design of Experiments (0 since not known in calcPPS)
2. Number of peaks
3. Number of peaks without LIP and RP
4. Reliable peaks (RP)
5. Peak picking score (PPS)

Author(s)

Gunnar Libiseller

See Also

`findIsotopes.IPO` `findIsotopes.CAMERA`

Examples

```
## Not run:
mzmlfile <- file.path(find.package("msdata"), "microtofq/MM14.mzML")
xset <- xcmsSet(mzmlfile, peakwidth=c(5,12), method="centWave")
calcPPS(xset)

## End(Not run)
```

`combineParams` *Combining two lists of parameters together.*

Description

This function combines two lists of parameters. The first is a list of parameters which should be optimized. This parameters have different values set by Design of Experimnt. The second list consists of parameters which should not be optimized, hence only one value is set for each parameter. The parameters of the second list are replicated to have the same length as the number experiments in the DoE. Then the two lists are combined.

Usage

```
combineParams(params_1, params_2)
```

Arguments

`params_1`	A list holding parameters which should be optimized. Each parameter already has value set for each experiment of an Design of Experiment.
`params_2`	A list holding parameters which should not be optimized, hence only one value is set.

Details

Special treatment is needed for the findPeaks.matchedFilter-parameters 'sigma', 'mzdiff' since these two parameters are defined by default relative to the parameters 'fwhm' or 'step' and 'steps' respectively.
`sigma=fwhm/2.3548` `mzdiff=0.8-step*steps`

Value

A list of consting of all parameters needed for an xcms-method (findPeaks.centWave, findPeaks.matchedFilter, retcor.obiwarp or group.density). Each list item has the same length which is equal to the number of experiments within the DoE.

Author(s)

Gunnar Libiseller

Examples

```
params <- getDefaultXcmsSetStartingParams()
typ_params <- typeCastParams(params)
design <- getBbdParameter(typ_params$to_optimize)
xcms_design <- decode.data(design)
xcms_design <- combineParams(xcms_design, typ_params$no_optimization)
xcms_design

## The function is currently defined as
function (params_1, params_2)
{
    len <- max(unlist(sapply(params_1, length)))
    p_names <- c(names(params_1), names(params_2))
    matchedFilter <- "fwhm" %in% p_names
    for (i in 1:length(params_2)) {
        new_index <- length(params_1) + 1
        fact <- params_2[[i]]
        params_1[[new_index]] <- fact
        if (matchedFilter) {
            if (p_names[new_index] == "sigma" && fact == 0) {
                if ("fwhm" %in% names(params_1)) {
                  params_1[[new_index]][1:len] <- params_1$fwhm/2.3548
                }
                else {
                  params_1[[new_index]][1:len] <- params_2$fwhm/2.3548
                }
            }
            else {
                if (p_names[new_index] == "mzdiff" && fact ==
                  0) {
                  if ("step" %in% names(params_1)) {
                    if ("steps" %in% names(params_1)) {
                      params_1[[new_index]][1:len] <- 0.8 - params_1$step *
                        params_1$steps
                    }
                    else {
                      params_1[[new_index]][1:len] <- 0.8 - params_1$step *
                        params_2$steps
                    }
                  }
                  else {
                    if ("steps" %in% names(params_1)) {
                      params_1[[new_index]][1:len] <- 0.8 - params_2$step *
                        params_1$steps
                    }
                    else {
                      params_1[[new_index]][1:len] <- 0.8 - params_2$step *
                        params_2$steps
                    }
                  }
                }
                else {
                  params_1[[new_index]][1:len] <- fact
                }
            }
```

```
            }
            else {
                params_1[[new_index]][1:len] <- fact
            }
        }
        names(params_1) <- p_names
        return(params_1)
    }
```

---

createModel *Creating a response surface model.*

Description

This function uses a design of experiments, a response for the experiments within the design and the used parameters to create a response surface model

Usage

```
createModel(design, params, resp)
```

Arguments

design	A design of experiments (Box-Behnken-Design or Central-Composite-Design)
params	The parameters which were used.
resp	The responses achivied for the various experiments.

Details

This function uses a design of experiments, a response for the experiments within the design and the used parameters to create a response surface model

Value

A response surface model.

Note

getBbdParameter getCcdParameter typeCastParams

Author(s)

Gunnar Libiseller

References

Lenth, R. V. (2009). Response-Surface Methods in R , Using rsm. Journal of Statistical Software, 32(7), 1-17. Retrieved from http://www.jstatsoft.org/v32/i07

Examples

```
params <- getDefaultXcmsSetStartingParams()
type_params <- typeCastParams(params)
design <- getBbdParameter(type_params$to_optimize)
resp <- runif(nrow(design),1,3)

model <- createModel(design, type_params$to_optimize, resp)
dev.new()
par(mfrow=c(3,2))
contour(model, ~ x1*x2*x3*x4, image=TRUE)
```

---

decode *En-/decodes values to/from ranges of -1 to 1.*

Description

Encode and decode values that are in a range of -1 to 1 into a specified range.

Usage

```
encode(value, bounds)
decode(value, bounds)
decodeAll(values, params)
```

Arguments

value	A value
values	A vector with values in the range [-1,1]
bounds	A vector of two values defining the lower and upper bound of a range.
params	A list where evere list-item consist of two values defining a lower and an upper bound.

Details

Decodes a values from ranges of -1 to 1 to ranges specified.

A function used to decode values that are in a range of -1 to 1 into a specified range. For every value a list item with lower and upper bound has to be supplied.

A function used to encode values that are in a specified range into a range between -1 to 1.

Value

decode: The encoded value. decodeAll: A vector of decoded values.

Author(s)

Gunnar Libiseller

Examples

```
decode(0, c(10, 20))
decode(-0.5, c(10, 20))
decode(1, c(10, 20))

bounds <- c(10, 20)
encode(decode(1, bounds), bounds)

        ## Multiple values:
values <- c(-1, -0.25, 0, 0.75)
params <- getDefaultXcmsSetStartingParams()
type_params <- typeCastParams(params)

decodeAll(values, type_params$to_optimize)

        ## Combination of encode and decode
encode(15, c(10, 20))
encode(10, c(10, 20))
encode(5, c(1, 5))

bounds <- c(1,5)
decode(encode(5, bounds), bounds)
```

`findIsotopes.CAMERA` *Identification of Isotopes using the package CAMERA.*

Description

This function finds isotopes using CAMERA's find peak function. Isotopes are separately found within each sample.

Usage

```
findIsotopes.CAMERA(xset, ...)
```

Arguments

`xset`	`xcmsSet` object
`...`	Additional parameters to the findIsotopes function of CAMERA

Details

Identification of 13C isotopes

Value

An matrix with 2 columns. Column one shows the peak id of the 12C, peak column two shows the id of the respective 13C isotope peak.

Author(s)

Gunnar Libiseller

References

C. Kuhl and R. Tautenhahn and C. Boettcher and T. R. Larson and S. Neumann: CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets Analytical Chemistry 84:283 (2012)

See Also

findIsotopes.IPO

Examples

```
## Not run:
mzmlfile <- file.path(find.package("msdata"), "microtofq/MM14.mzML")
xset <- xcmsSet(mzmlfile, peakwidth=c(5,12), method="centWave")
isotopes <- findIsotopes.CAMERA(xset, ppm=15, maxcharge=2)

## End(Not run)
```

---

findIsotopes.IPO *Identification of 13C isotopes*

Description

This function identifies natural, stable 13C isotopes within an xcmsSet object of LC-HRMS data. Isotopes have to be within a mass-, retentiontime- and intensity- window to be recognized as isotopes. If checkBorderIntensity is TRUE the maximum intensity of each peaks has to be at least three times the intensity at rtmin and rtmax.

Usage

```
findIsotopes.IPO(xset, checkBorderIntensity=FALSE)
```

Arguments

xset	xcmsSet object
checkBorderIntensity	logical whether the peaks 'maxo' value has to be at least three times the intensities at 'rtmin' and 'rtmax'

Details

Identification of 13C isotopes

Value

An matrix with 2 columns. Column one shows the peak id of the 12C, peak column two shows the id of the respective 13C isotope peak.

Author(s)

Gunnar Libiseller

See Also

findIsotopes.CAMERA

Examples

```
## Not run:
mzmlfile <- file.path(find.package("msdata"), "microtofq/MM14.mzML")
xset <- xcmsSet(mzmlfile, peakwidth=c(5,12), method="centWave")
isotopes <- findIsotopes.IPO(xset, TRUE)

## End(Not run)
```

---

getBbdParameter *Creates a Box-Behnken Design of experiment*

Description

Creates a Box-Behnken Design of Experiment out of a list of parameters. Each of the list items has to be a pair defining the lower und upper limits of the value-range to test. The method then returns a Center faced Box-Behnken Design of Experiments. The list has to hold a least three pairs.

Usage

```
getBbdParameter(params)
```

Arguments

params	A list of value pairs defining lower und upper limits of an optimization range.

Details

Creates a Box-Behnken Design of Experiment out of a list of parameters. Each of the list items has to be a pair defining the lower und upper limits of the value-range to test. The method then returns a Center faced Box-Behnken Design of Experiments. The list has to hold a least three pairs.

Value

A Box-Behnken Design of Experiments

Author(s)

Gunnar Libiseller

References

Lenth, R. V. (2009). Response-Surface Methods in R , Using rsm. Journal of Statistical Software, 32(7), 1-17. Retrieved from http://www.jstatsoft.org/v32/i07

See Also

getCcdParameter

Examples

```
params <- getDefaultXcmsSetStartingParams()
typ_params <- typeCastParams(params)
design <- getBbdParameter(typ_params$to_optimize)

## The function is currently defined as
function (params)
{
    require(rsm)
    lower_bounds <- unlist(lapply(X = params, FUN = function(x) x[1]))
    higher_bounds <- unlist(lapply(X = params, FUN = function(x) x[2]))
    steps <- (higher_bounds - lower_bounds)/2
    x <- paste("x", 1:length(params), " ~ (", c(names(params)),
        " - ", (lower_bounds + steps), ")/", steps, sep = "")
    formulae <- list()
    for (i in 1:length(x)) formulae[[i]] <- as.formula(x[i])
    design <- bbd(length(params), n0 = 1, randomize = FALSE,
        coding = formulae)
    return(design)
  }
```

`getCcdParameter` *Creates a Central-Composite Design of experiment*

Description

Creates a Central-Composite Design of Experiment out of a list of parameters. Each of the list items has to be a pair defining the lower und upper limits of the value-range to test. The method then returns a Center faced Central-Composite Design of Experiments.

Usage

```
getCcdParameter(params)
```

Arguments

`params`	A list of value pairs defining lower und upper limits of an optimization range.

Details

Creates a Central-Composite Design of Experiment out of a list of parameters. Each of the list items has to be a pair defining the lower und upper limits of the value-range to test. The method then returns a Center faced Central-Composite Design of Experiments.

Value

A Central-Composite Design of Experiments

Author(s)

Gunnar Libiseller

References

Lenth, R. V. (2009). Response-Surface Methods in R , Using rsm. Journal of Statistical Software, 32(7), 1-17. Retrieved from http://www.jstatsoft.org/v32/i07

See Also

getBbdParameter

Examples

```
params <- getDefaultXcmsSetStartingParams()
typ_params <- typeCastParams(params)
design <- getCcdParameter(typ_params$to_optimize)

## The function is currently defined as
function (params)
{
    require(rsm)
    lower_bounds <- unlist(lapply(X = params, FUN = function(x) x[1]))
    higher_bounds <- unlist(lapply(X = params, FUN = function(x) x[2]))
    steps <- (higher_bounds - lower_bounds)/2
    x <- paste("x", 1:length(params), " ~ (", c(names(params)),
        " - ", (lower_bounds + steps), ")/", steps, sep = "")
    formulae <- list()
    for (i in 1:length(x)) formulae[[i]] <- as.formula(x[i])
    design <- ccd(length(params), n0 = 1, alpha = "face", randomize = FALSE,
        inscribed = TRUE, coding = formulae)
    return(design)
  }
```

---

getDefaultRetCorCenterSample

Gets the index of the sample with most peaks in it.

Description

Gets the index of the sample with most peaks in it. This is used if no center sample for retention time correction has been defined by the user.

Usage

```
getDefaultRetCorCenterSample(xset)
```

Arguments

xset	xcmsSet object

Details

Gets the index of the sample with most peaks in it. This is used if no center sample for retention time correction has been defined by the user.

Value

The file index of the sample with most peaks in it.

Author(s)

Gunnar Libiseller

Examples

```
## The function is currently defined as
function (xset)
{
    ret <- NULL
    for (i in 1:length(xset@filepaths)) {
        ret <- c(ret, sum(xset@peaks[, "sample"] == i))
    }
    return(which.max(ret))
  }
```

`getDefaultRetGroupStartingParams` *Gives a List of parameters for xcms-methods retcor.obiwarp or retcor.loess and group.density which are optimized by default*

Description

This function creates a list of parameters used in the xcms-methods retcor.obiwarp and group.density.
Per default the following parameters have a defined range where optimization should start:
retcor.obiwarp parameters: 'gapInit'; 'gapExtend', 'profStep'
group.density parameters: 'bw', 'minfrac', 'mzwid'

Usage

```
getDefaultRetGroupStartingParams(retcorMethod=c("obiwarp", "loess", "none"), distfunc=c("cor_op
```

Arguments

`retcorMethod`	The name of the retention time correction method that should be used. The XCMS methods retcor.obiwarp and retcor.loess are supported. If no retention time correction should be done use "none".
`distfunc`	The name of the distance function used by retcor.obiwarp
`high_resolution`	If high_resolution = TRUE starting values for mzwid are set to 0.015 and 0.035; if high_resolution = FALSE to 0.15, 0.35

Details

* Do not delete a parameter from the list returned.
* Optimization of qualitative parameters is not supported yet.
* If you want to optimize additional parameter just set an lower and an upper bound (e.g. params$max <- c(4,8))
* If you dont want to optimize a parameter set a default value (e.g. params$max <- 10)

Value

A List of parameters used in the xcms-methods retcor.obiwarp or retcor.loess and group.density

Author(s)

Gunnar Libiseller

Examples

```
params <- getDefaultRetGroupStartingParams()
params$bw <- 10
params$max <- c(4,8)
params
```

`getDefaultXcmsSetStartingParams` *Creates a List of parameters for xcms-methods xcmsSet.findPeak which are optimized by default*

Description

This function creates a list of parameters used in the xcmsSet.findPeak-methods 'centWave' and 'matchedFilter'. Per default the following parameters have a defined range where optimization should start:
'centWave' parameters: 'peakwidth' (split into 'min_peakwidth' and 'max_peakwidth'), 'ppm', 'mzdiff'
'matchedFilter' parameters: 'fwhm', 'snthresh', 'step', 'steps'

Usage

```
getDefaultXcmsSetStartingParams(method = c("centWave", "matchedFilter"))
```

Arguments

`method`	Either parameters for 'centWave' or 'matchedFilter' should be created

Details

* Do not delete a parameter from the list returned.
* Optimization of qualitative parameters is not supported yet.
* If you want to optimize additional parameter just set an lower and an upper bound (e.g. params$snthresh <- c(5,20))
* If you dont want to optimize a parameter set a default value (e.g. params$snthresh <- 10)

Value

A List of parameters for the xcmsSet.findPeak-methods 'centWave' or 'matchedFilter'

Author(s)

Gunnar Libiseller

Examples

```
params <- getDefaultXcmsSetStartingParams()
params$ppm <- 10
params$snthresh <- c(5,15)
params


## The function is currently defined as
function(method=c("centWave", "matchedFilter")) {

    method <- match.arg(method)

    if(method=="centWave")
      return(list(min_peakwidth=c(12,28), max_peakwidth=c(35,65), ppm=c(17,32),
                  mzdiff=c(-0.001, 0.01), snthresh=10, noise=0, prefilter=3,
                  value_of_prefilter=100,  mzCenterFun="wMean", integrate=1,
                  fitgauss=FALSE, verbose.columns=FALSE))

    if(method=="matchedFilter")
      return(list(fwhm=c(25,35), snthresh=c(3,17), step=c(0.05, 0.15), steps=c(1,3),
                  sigma=0, max=5, mzdiff=0, index=FALSE))

  }
```

---

getNormalizedResponse *It combines Retention time Correction Scores (RCS) and Grouping Scores (GS)*

Description

This function does unity based normalization on Retention time Correction Scores (RCS) as well as Grouping Scores (GS).

Usage

```
getNormalizedResponse(response)
```

Arguments

response	A List of all responses calculated by getRGTVValues for all experiments of an Design of Experiment

Details

Grouping Score (GS) is calculated by:
'good groups'^2/'bad groups

For all RCS and GS values unitiy based normalization is done. For every experiment within the DoE these two values are added together and returned.

Value

A vector with RTGV values

Note

Since RCS and GS can be within completely different ranges, normalization has to be done to prevent an excessive influence of either RCS or GS.

Author(s)

Gunnar Libiseller

See Also

getRGTVValues

Examples

```
##---- Should be DIRECTLY executable !! ----
##-- ==>  Define data, use random,
##--or do  help(data=index)  for the standard data sets.

## The function is currently defined as
function (response)
{
    good_groups <- sapply(response, "[[", "good_groups")
    bad_groups <- sapply(response, "[[", "bad_groups")
    bad_groups_bool <- bad_groups == 0
    bad_groups[bad_groups_bool] <- 1
    good_groups[bad_groups_bool] <- good_groups[bad_groups_bool] +
        1
    group_ratio <- good_groups^2/bad_groups
    ARTS <- 1/sapply(response, "[[", "mean_rel_rt_diff")
    ARTS_penalty <- 1/sapply(response, "[[", "retcor_done")
    ARTS <- ARTS/ARTS_penalty
    norm_group_ratio <- (group_ratio - min(group_ratio))/(max(group_ratio) -
        min(group_ratio))
    norm_ARTS <- (ARTS - min(ARTS))/(max(ARTS) - min(ARTS))
    return(norm_group_ratio + norm_ARTS)
  }
```

`getRGTVValues` *Calculation of values used for a Retention time correction and Grouping Target Value (RGTV)*

Description

This function calculates the Retention time Correction Score (RCS) of all features within an xcmsSet-object. Also features having exactly one peak from each sample are defined as 'good groups', all others a 'bad groups'.

Usage

```
getRGTVValues(xset, exp_index = 1, retcor_penalty = 1)
```

Arguments

`xset`	`xcmsSet` object
`exp_index`	Experiment-id of the experiment within a Design of Experiments
`retcor_penalty`	Penalty if an error occured with the used retention time correction parameters

Details

This function calculates the Retention time Correction Score (RCS) of all features within an xcmsSet-object. Also features having exactly one peak from each sample are defined as 'good groups', all others a 'bad groups' which leads to a Grouping Score (GS) by calculating 'good groups'^2/'bad groups'.

Value

a list containing the items exp_index, good_groups, bad_groups, GS and RCS.

Author(s)

Gunnar Libiseller

Examples

```
## Not run:
getRGTVValues(xset)

## End(Not run)
```

`optimizeRetGroup` *Optimization for parameters for retention time correction and grouping*

Description

This function provides optimisation for parameters of the xcms-method retcor.obiwarp and group.density. The retention time correction is optimised by minimizing intra-feature retention time shifts; grouping is optimized by increasing the number of features which have exactly one peak per sample.

Usage

```
optimizeRetGroup(xset, params = getDefaultRetGroupStartingParams(),
                 nSlaves = 4, subdir = "IPO")
```

Arguments

`xset`	`xcmsSet` object
`params`	A list of parameters which are needed by xcms-methods retcor.obiwarp and group.density. List-items with two values will be optimized. The first value defines the lower test value, the second one the higher test value.
`nSlaves`	Number of slaves the optimization process should spawn.
`subdir`	The name of the subdirectory which is created and where the figures of the response surface models will be saved to.

Details

This function provides optimisation for parameters of the xcms-method retcor.obiwarp and group.density. The retention time correction is optimised by minimizing intra-feature retention time shifts; grouping is optimized by increasing the number of features which have exactly one peak per sample.

Value

A LIST of length n+1 with n beeing the optimization runs needed

`comp1-comp(n)`	A LIST containing: 1. Parameters used for the nth optimization run 2. Box-Behnken or Central Composite Design used for optimization run 3. Responses from calculateRGTV for every experiment in the design 4. Response surface model for the design 5. The normalized parameter settings giving the best Retention time and Grouping Target Value (RGTV) (values between -1 and 1) 6. Response from calculateRGTV for xcmsSet-object created with best parameters in this run 7. xcmsSet-object created with best parameters in this run
`comp(n+1)`	A LIST containing: 1. Parameters giving the best RGTV

Author(s)

Gunnar Libiseller

References

Obiwarp Prince, J. T., & Marcotte, E. M. (2006). Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Analytical chemistry, 78(17), 6140-52. doi:10.1021/ac0605344

See Also

getDefaultRetGroupStartingParams

Examples

```
## Not run:

library(IPO)

# use this to get an xcmsSet-object with optimized settings or any other xset
mzdatafiles <- list.files(file.path(find.package("mtbls2"), "mzdata"),
                          full.names=TRUE)[1:4]
paramsPP <- getDefaultXcmsSetStartingParams()
paramsPP$mzdiff <- -0.001
paramsPP$min_peakwidth <- c(7,14)
paramsPP$max_peakwidth <- c(20,30)
resultPP <- optimizeXcmsSet(mzdatafiles, paramsPP, nSlaves=4, subdir="mtbls2")

#optimize the retention time correction and grouping parameters
paramsRG <- getDefaultRetGroupStartingParams()
paramsRG$profStep <- 1
paramsRG$minfrac <- 0.75
resultRG <- optimizeRetGroup(resultPP$best_settings$xset, params=paramsRG,
                             nSlaves=4,subdir="mtbls2")

writeRScript(resultPP$best_settings$parameters, resultRG$best_settings, 4)


## End(Not run)
```

`optimizeXcmsSet` *Optimisation of peak picking parameters by using natural, stable 13C isotopes*

Description

This function provides optimisation of peak picking parameters by using natural, stable 13C isotopes.

Usage

```
    optimizeXcmsSet(files, params = getDefaultXcmsSetStartingParams(),
isotopeIdentification=c("IPO", "CAMERA"), nSlaves = 4, subdir = "IPO", ...)
```

Arguments

files	A directory or list of files, passed to xcmsSet(). If no files are given, xcmsSet() will check recursively all MS files in the current working directory.
params	A list of parameters which are needed by XCMS::findPeaks-Methods. List-items with two values will be optimized. The first value defines the lower test value, the second one the higher test value.
isotopeIdentification	This parameter defines the method for isotope identification. The method 'IPO' was especially implemented for high resolution data. CAMERA is an established isotope and adduct annotation package.
nSlaves	Number of slaves the optimization process should spawn.
subdir	The name of the subdirectory which is created and where the figures of the response surface models will be saved to.
...	Additional parameters to CAMERA's or IPO's findIsotopes functions

Details

This function provides optimisation of peak picking parameters by using natural, stable 13C isotopes.

Value

A LIST of length n+1 with n beeing the optimization runs needed

comp1-comp(n)	A LIST containing: 1. Parameters used for the nth optimization run 2. Box-Behnken or Central Composite Design used for optimization run 3. Responses from calcPPS for every experiment in the design 4. Response surface model for the design 5. The normalized parameter settings giving the best PPS (values between -1 and 1)
comp(n+1)	A LIST containing: 1. Parameters giving the best PPS 2. An xcmsSet-object created with the optimized parameters 3. The result of calcPPS() of the xcmsSet created with the best parameters

Author(s)

Gunnar Libiseller

References

Smith, C.A. and Want, E.J. and O'Maille, G. and Abagyan,R. and Siuzdak, G.: XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification, Analytical Chemistry, 78:779-787 (2006)

Ralf Tautenhahn, Christoph Boettcher, Steffen Neumann: Highly sensitive feature detection for high resolution LC/MS BMC Bioinformatics, 9:504 (2008)

H. Paul Benton, Elizabeth J. Want and Timothy M. D. Ebbels: Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data Bioinformatics, 26:2488 (2010)

C. Kuhl and R. Tautenhahn and C. Boettcher and T. R. Larson and S. Neumann: CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets Analytical Chemistry 84:283 (2012)

See Also

getDefaultXcmsSetStartingParams calcPPS findIsotopes.IPO findIsotopes.CAMERA

Examples

```
## Not run:
library(IPO)

mzdatafiles <- list.files(file.path(find.package("mtbls2"), "mzdata"),
                          full.names=TRUE)[1:4]

paramsPP <- getDefaultXcmsSetStartingParams()
paramsPP$mzdiff <- -0.001
paramsPP$min_peakwidth <- c(7,14)
paramsPP$max_peakwidth <- c(20,30)
#example using IPO isotope identification
resultPP <- optimizeXcmsSet(mzdatafiles, paramsPP, subdir="mtbls2")

#example using CAMERA isotope identification
resultPP <- optimizeXcmsSet(mzdatafiles, paramsPP, isotopeIdentification="CAMERA",
                            subdir="mtbls2", ppm=15, maxcharge=2)


## End(Not run)
```

---

toMatrix *Converts an array into a matrix*

Description

This function converts an array into a matrix. This is useful to counter the implicit casting of matrices into arrays when only one row is selected. If a matrix is passed to the function, the matrix is returned, if an array is passed, a matrix with one row is returned.

Usage

```
toMatrix(data)
```

Arguments

data	An array or a matrix

Value

A matrix

Author(s)

Gunnar Libsieller

Examples

```
data <- matrix(1:9, nrow=3)
colnames(data) <- c("a","b","c")
x <- data[1,]
is.matrix(x)
x <- toMatrix(x)
is.matrix(x)
```

`typeCastParams` *Splits parameters into those which should be optimized and those which are fixed.*

Description

This method takes a list of parameters and returns a list consisting of another two lists; one holding parameters ment for optimization and one holding fixed parameters.

Usage

```
typeCastParams(params)
```

Arguments

`params`	A list of parameters for an xcms-method

Details

This method takes a list of parameters and returns a list consisting of another two lists; one holding parameters ment for optimization and one holding fixed parameters.

Value

A list of:

`to_optimize`	A LIST containing all parameters which should be optimized.
`no_optimization`	A LIST containing all parameters which should not be optimized.

Author(s)

Gunnar Libiseller

See Also

`optimizeXcmsSet`, `optimizeRetGroup`

Examples

```
## The function is currently defined as
function (params)
{
    ret_1 <- list()
    ret_2 <- list()
    ret <- list()
    for (i in 1:length(params)) {
        factor <- params[[i]]
        if (length(factor) == 2) {
            ret_1[[(length(ret_1) + 1)]] <- factor
            names(ret_1)[length(ret_1)] <- names(params)[i]
        }
        else {
            ret_2[[(length(ret_2) + 1)]] <- factor
            names(ret_2)[length(ret_2)] <- names(params)[i]
        }
    }
    ret$to_optimize <- ret_1
    ret$no_optimization <- ret_2
    return(ret)
  }
```

`writeParamsTable` *Writes XCMS settings to a file.*

Description

This function write findPeaks, retcor and grouping parameters to a file using write.table.

Usage

```
writeParamsTable(peakPickingSettings, retCorGroupSettings, file, ...)
```

Arguments

`peakPickingSettings`	A list of optimized settings for xcms-methods findPeaks.centWave or findPeaks.matchedFilter
`retCorGroupSettings`	A list of optimized settings for xcms-methods for retcor.obiwarp and group.density
`file`	The name of the outputfile for the parameters.
`...`	Additional parameters for write.table.

Details

This function write findPeaks, retcor and grouping parameters to a file using write.table.

Value

none

Author(s)

Gunnar Libiseller

See Also

xcms

---

writeRScript *Prints an R-script to the screen which can be used for xcms processing*

Description

This function prints a script of the optimized findPeaks, retcor and grouping parameters to the screen.

Usage

```
writeRScript(peakPickingSettings, retCorGroupSettings, nSlaves)
```

Arguments

peakPickingSettings	The optimized settings for xcms-methods findPeaks.centWave or findPeaks.matchedFilter
retCorGroupSettings	The optimized settings for xcms-methods for retcor.obiwarp and group.density
nSlaves	The number of slaves which should be used in the script

Details

This function that prints a script out of the optimized findPeaks, retcor and grouping parameters to the screen.

Value

none

Author(s)

Gunnar Libiseller

Index