Thesis for the degree of Doctor of Technology, Sundsvall 2010

EFFICIENT ALGORITHMS FOR HIGHLY AUTOMATED EVALUATION OF LIQUID CHROMATOGRAPHY - MASS SPECTROMETRY DATA

Mattias Fredriksson

Supervisors:
Dan Bylund
Patrik Petersson
Bengt-Olof Axelsson

Department of Natural Sciences, Engineering and Mathematics
Mid Sweden University, SE-851 70 Sundsvall, Sweden

ISSN 1652-893X,
Mid Sweden University Doctoral Thesis 98
ISBN 978-91-86694-03-6

Akademisk avhandling som med tillstånd av Mittuniversitetet i Sundsvall framläggs till offentlig granskning för avläggande av teknologie doktorsexamen, fredag 3 december 2010, klockan 10:15 i sal L111, Mittuniversitetet, Sundsvall. Seminariet kommer att hållas på svenska.

EFFICIENT ALGORITHMS FOR HIGHLY AUTOMATED EVALUATION OF LIQUID CHROMATOGRAPHY - MASS SPECTROMETRY DATA

Mattias Fredriksson

Department of Natural Sciences, Engineering and Mathematics
Mid Sweden University, SE-851 70 Sundsvall
Sweden

Telephone: +46 (0)771-975 000

Printed by Kopieringen Mid Sweden University, Sundsvall, Sweden, 2010

EFFICIENT ALGORITHMS FOR HIGHLY AUTOMATED EVALUATION OF LIQUID CHROMATOGRAPHY - MASS SPECTROMETRY DATA

Mattias Fredriksson

Department of Natural Sciences, Engineering and Mathematics
Mid Sweden University, SE-851 70 Sundsvall, Sweden
ISSN 1652-893X, Mid Sweden University Doctoral Thesis 98; ISBN 978-91-86694-03-6

ABSTRACT

Liquid chromatography coupled to mass spectrometry (LC-MS) has due to its superior resolving capabilities become one of the most common analytical instruments for determining the constituents in an unknown sample. Each type of sample requires a specific set-up of the instrument parameters, a procedure referred to as method development. During the requisite experiments, a huge amount of data is acquired which often need to be scrutinised in several different ways. This thesis elucidates data processing methods for handling this type of data in an automated fashion.

The properties of different commonly used digital filters were compared for LC-MS data de-noising, of which one was later selected as an essential data processing step during a developed peak detection step. Reconstructed data was further discriminated into clusters with equal retention times into components by an adopted method. This enabled an unsupervised and accurate comparison and matching routine by which components from the same sample could be tracked during different chromatographic conditions.

The results show that the characteristics of the noise have an impact on the performance of the tested digital filters. Peak detection with the proposed method was robust to the tested noise and baseline variations but functioned optimally when the analytical peaks had a frequency band different from the uninformative parts of the signal. The algorithm could easily be tuned to handle adjacent peaks with lower resolution. It was possible to assign peaks into components without typical rotational and intensity ambiguities associated to common curve resolution methods, which are an alternative approach. The underlying functions for matching components between different experiments yielded satisfactory results. The methods have been tested on various experimental data with a high success rate.

Keywords: Digital filtering, Liquid chromatography, Mass spectrometry, Method development, Peak detection, Peak purity, Peak tracking

SAMMANFATTNING

De analysinstrument som används för att ta reda på vad ett prov innehåller (och till vilken mängd) måste vanligtvis ställas in för det specifika fallet, för att fungera optimalt. Det finns ofta en mängd olika variabler att undersöka som har mer eller mindre inverkan på resultatet och när provet är okänt kan man oftast inte förutspå de optimala inställningarna i förtid.

En vätskekromatograf med en masspektrometer som detektor är ett sådant instrument som är utvecklat för att separera och identifiera organiska ämnen lösta i vätska. Med detta mycket potenta system kan man ofta med rätt inställningar dela upp de ingående ämnena i provet var för sig och samtidigt erhålla mått som kan relateras till dess massa och mängd. Detta system används flitigt av analytiska laboratorer inom bl.a. läkemedelsindustrin för att undersöka stabilitet och renhet hos potentiella läkemedel. För att optimera instrumentet för det okända provet krävs dock att en hel del försök utförs där inställningarna varieras. Syftet är att med en mindre mängd designade försök bygga en modell som klarar av att peka åt vilket håll de optimala inställningarna finns. Data som genereras från instrumentet för denna typ av applikation är i matrisform då instrumentet scannar och sparar intensiteten av ett intervall av massor varje tidpunkt en mätning sker. Om en analyt når detektorn vid aktuell tidpunkt återges det som en eller flera överlagda normalfördelade toppar som ett specifikt mönster på en annars oregelbunden bakgrundssignal. Förutom att alla topparna i det färdiga datasetet helst ska vara välseparerade och ha den rätta formen, så ska tiden analysen pågår vara så kort som möjlig. Det är ändå inte ovanligt att ett färdigt dataset består av tiotals miljoner uppmätta intensiteter och att det kan krävas runt 10 försök med olika betingelser för att åstadkomma ett godtagbart resultat.

Dataseten kan dock till mycket stor del innehålla brus och andra störande signaler vilket gör de extra krångligt att tolka och utvärdera. Eftersom man även ofta får att komponenterna byter plats i ett dataset när betingelserna ändras kan en manuell utvärdering ta mycket lång tid.

Syftet med denna avhandling har varit att hitta metoder som kan vara till nytta för den som snabbt och automatiskt behöver jämföra dataset analyserade med olika kromatografiska betingelser, men med samma prov. Det slutgiltiga målet har främst varit att identifiera hur olika komponenter i provet har rört sig mellan de olika dataseten, men de steg som ingår kan även nyttjas till andra applikationer.

TABLE OF CONTENTS

LIST OF PAPERS

This thesis is mainly based on the following four papers, herein referred to by their Roman numerals:

Paper I **An objective comparison of pre-processing methods for enhancement of liquid chromatography - mass spectrometry data**
Mattias Fredriksson, Patrik Petersson, Magnus Jörntén-Karlsson, Bengt-Olof Axelsson, Dan Bylund
Journal of Chromatography A, 1172 (2007) 135–150

Paper II **An automatic peak finding method for liquid chromatography-mass spectrometry data using Gaussian second derivative filtering**
Mattias Fredriksson, Patrik Petersson, Bengt-Olof Axelsson, Dan Bylund
Journal of Separation Science, 32 (2009) 3906–3918

Paper III **A component tracking algorithm for accelerated and improved liquid chromatography - mass spectrometry method development**
Mattias Fredriksson, Patrik Petersson, Bengt-Olof Axelsson, Dan Bylund
Accepted for publication in Journal of Chromatography A

Paper IV **Combined use of algorithms for peak picking, peak tracking and retention modelling to optimize the chromatographic conditions for liquid chromatography - mass spectrometry analysis of fluocinolone acetonide and its degradation products**
Mattias Fredriksson, Patrik Petersson, Bengt-Olof Axelsson, Dan Bylund
Submitted to Analytica Chimica Acta

1. INTRODUCTION

Most commercial pharmaceutical drugs have an expiration date. Beyond this time period, the safety of the product can no longer be guaranteed. Retrieving the required background data to settle the long time storage date requires a great deal of both effort and time. The methods discussed in this thesis can be employed to reduce both.

Before a new drug substance is released to the market, all possible hazardous components, associated to the degradation of the active pharmaceutical ingredient, side products, solvent residues from manufacturing or leachables from the packaging need to be detected, determined and evaluated. A commonly used and highly sensitive and selective instrument combination for analysing these kinds of substances is liquid chromatography coupled to mass spectrometry (LC-MS). The drug substances are then dissolved and introduced to the instrument where the sample constituents are separated and detected. The resulting data set can be seen as an ocean of noise, in which the sample constituents have made patterns in form of peaks more or less resolved in time and mass. The information gained from the patterns is used to identify and quantify the sample components. To be able to decipher the patterns as clearly as possible, they have to be acceptably separated from each other and conspicuous. In a sample with unknown constituents, it is difficult to predict in advance the optimal instrumental set-up. Therefore the analyst tests the same sample several times using different instrument conditions in a procedure referred to as method development. The patterns can then arise at completely different positions in the data. This generates a new problem formulation; the peaks have to be identified and tracked in all data sets, commonly a difficult and tedious manual work even for an experienced analytical chemist due to the huge amount of data generated.

The annoyance over this bottleneck was the major driving force to this thesis work. Could the tracking of the sample components between different analytical runs be performed automatically or at least semi-automatically to increase analysis throughput? Several difficult passages had to be defeated for this to be feasible. Increasing the possibility of finding the relevant peaks in each data set was a natural first step, and a selection of methods was evaluated in *paper I*. The most appropriate method was then further developed to be able to fully automatically detect and highlight the peaks and remove the noise constituents. Measuring the multi-facetted noise in a decent manner and automating the process were two hurdles that had to be overcome as described in *paper II*. The resulting peaks then had to be assigned to the correct component; this is a rather easy procedure when

components are decently separated from each other in time, but rather difficult when they are not. Furthermore, when all data sets of interest had their peak patterns assigned into components they could finally be tracked between the different data sets by a similarity measurement, a procedure described in *paper III*. The proposed strategy was finally applied to the last step of the method development in *paper IV*, where components present in several data sets needed to be tracked.

The following pages cover some of the techniques used to enhance, refine, extract and unscramble LC-MS data to aid the methodological progress based on the papers discussed. The text includes an introduction to the instruments used and a complete strategy for improving efficiency during LC-MS method development for a typical pharmaceutical drug and its degradation products. The main focus of the thesis is the explanation of the strategies used to reach the established goals, along with their benefits and limitations.

2. THE ANALYTICAL INSTRUMENTS

The endeavour of separating the molecules present in a sample mixture has been under development since the first successful attempts during the first decade of the 20th century. The aim is often to identify, quantify or purify the individual components in the sample and various techniques has evolved to handle almost any type of mixtures. Two relevant analytical instruments for determining the composition of an unknown sample mixture and the level of the constituents are the liquid chromatograph and the mass spectrometer. These possess both qualitative and quantitative properties.

2.1. Liquid chromatography

Reversed phase high performance liquid chromatography (RP-HPLC, or just LC) is a common technique for separating organic molecules in a sample. The sample is injected to a system where one or more pumps continuously deliver a polar mobile phase (often purified water) with an organic modifier through tubing into a column. The column contains a packing material that has roughly the same polarity as the components to be separated. All components should, however, have slightly different affinities for the material for optimal performance. The components in the sample then elute, one at the time optimally, from the column with the mobile phase and are then further detected by one of several methods. An organic modifier, commonly methanol or acetonitrile, is used to change the

velocity of the components through the column since the organic molecules in the sample then obtain a greater affinity to the mobile phase compared to pure water. RP-HPLC has showed very good results in separating many non-volatile organic substances and is by far the most common instrument in most analytical labs within the pharmaceutical industry.

To detect the separated components eluting from the column some kind of detector is needed. The most common detector today is ultra-violet (UV) detection where the sample is lit by a lamp that emits UV radiation. The absorbance of the molecules is registered when they pass the lamp in a flow cell. Nowadays, the UV detectors can register the absorbance of several wavelengths simultaneously by so-called diode array detectors (DAD).Since different chemical groups have more or less different absorbance spectra, the sample components can be differentiated by more than intensity alone. The detector measures the amount eluting in real time. Since band broadening occurs in the system, mainly due to diffusion, the recorded presence of a sample component will be a bell shaped (Gaussian) peak. The resulting recording is referred to as a chromatogram.

There are several parameters that influence the retention time, selectiveness, resolution and peak shape of the sample components in the column. Column parameters (length, inner diameter, type of packing material and temperature) and mobile phase parameters (type of organic modifier, buffer and pH) are the most commonly optimised. Some of the parameters have a greater impact than others.

During isocratic analysis, the parameters are kept constant throughout the analysis, which generally results in smaller peak widths in the beginning of the chromatogram and broader peak widths in the end due to an increased diffusion in the column with longer duration of stay. Peak shapes can be preserved by using a gradient system where the proportion of organic modifier is continuously increased as analysis progress. This pushes the sample components with a higher affinity to the column through the system faster, which reduces the effect of diffusion. Different sample constituents can thus be eluted within reasonable analysis time. Inside the column, the elution rate increases in the tail compared to the front, which compresses the peak. For some specific peak widths, the dispersion and compression cancels out. The mechanisms behind the retention behaviour are thoroughly investigated and can be modelled rather accurately [1].

The development of liquid chromatography is moving towards even shorter analysis times through the use of systems that manage the higher backpressures associated with the use of efficient columns packed with sub-2 μm particles or sub-3 μm superficially porous particles (fused-core) [2,3].

2.2. Mass spectrometry

The mass spectrometer is a very sensitive and specific instrument that is capable of measuring the mass to charge (m/z) ratios and the amounts of ionisable molecules in a sample. This instrument is highly sophisticated and suitable for both qualitative and quantitative analysis. The sample is infused into the ion source where it becomes ionized (charged) before entering the mass analyser, which is operated under vacuum to avoid interference from molecules originating from the ambient air. Here, the m/z ratio is measured by one of several methods.

If the instrument is equipped with a quadrupole, a combination of AC and DC voltages is applied over four metal rods where the electricity is tuned such that only ions with a selected m/z ratio can pass the rods and hit the detector. An alternative and common mass analyser is the time-of-flight (TOF), where the m/z ratio is determined by measuring the time it takes for an ion to reach the detector after being exposed to an electric field of known strength. Ion trap is another mass analyser that can capture the ions by electric or magnetic fields where they can be manipulated to reveal their mass. All mass analysers have their respective advantages and disadvantages, but all attempt to produce a mass spectrum of all masses of the components in the sample.

2.3. Hyphenated LC-MS

The liquid chromatograph can be coupled to the mass spectrometer, which then serves as a detector. This generates a very potent system capable of separating the sample constituents both in time and mass by registering their relative elution time and measuring their m/z ratio one component at a time. This system has the advantage over LC-UV that the obtained spectra are more specific. Since the UV detector does not destroy the sample, the UV detector can be incorporated so that a LC-UV-MS system is obtained. In this manner, only components that are both non-ionisable and at the same time lacking in chromophoric groups will remain undetected. In **Fig 1**, a schematic figure of the hyphenated LC-UV-MS is shown.

2.3.1 Electrospray ionisation

The interface between the end of the LC system and the inlet of the mass spectrometer, where the sample must be converted from the liquid phase and atmospheric pressure, to the particle phase and vacuum was for a long time difficult to obtain. The electrospray ionisation (ESI) chamber, however, is a soft

ionisation technique that manages to keep the molecules intact in most cases and is often the best choice when analysis of larger molecules is required.

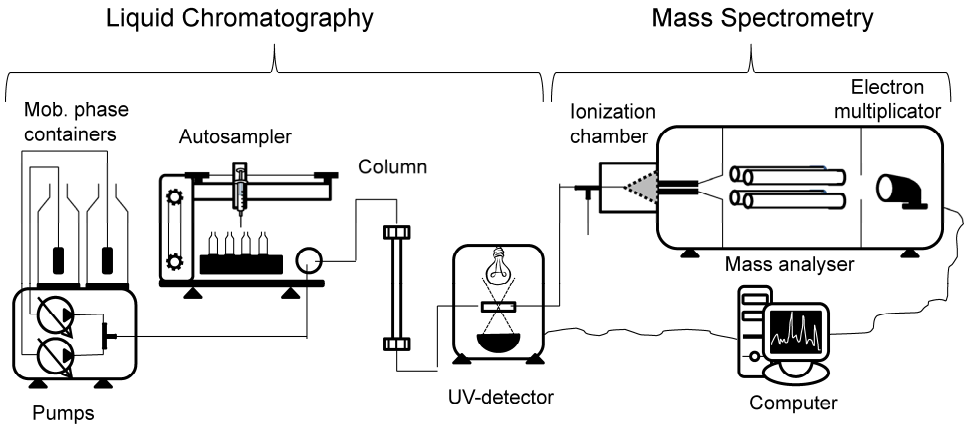

Figure 1. Schematic figure of a hyphenated LC-UV-MS system with possibility to use gradient elution.

The analytes travel with the mobile phase into the electrospray ionization chamber, where the liquid is charged by an applied potential that generates an electric field between the outlet needle at the end of the tubing, and the MS inlet. Electro-chemical reactions then occur which leads to an excess of charges in the solution. If the potential is high enough for the current mobile phase composition, the liquid forms a Taylor cone whereby charged droplets are formed when the columbic repulsion exceeds the surface tension (i.e. Rayleigh limit). The droplet formation is supported by a nebulising gas and a drying gas can be used to assist in the evaporation of the solvent (i.e. mobile phase). Further evaporation generates an increasingly higher concentration of charges in the droplets, which makes them unstable due to the repulsive forces. In the closing stages the droplets practically explode, a process known as columb fission that renders smaller droplets with even higher charge density. This process is repeated until single ions remain, which are then guided into the mass spectrometer by the vacuum and by the electric fields applied. A schematic picture of the ESI chamber is shown in **Fig. 2**.

There are some limitations associated with ESI-MS that render the resulting data ambiguous. The obtained m/z ratios may not directly correspond to the mass of the sample components. Larger molecules can obtain several charges that then decrease the apparent m/z ratio as many times as the number of charges, ions can form cluster molecules with itself (dimers) or components from the buffer in the

mobile phase (adducts), or the molecules can break apart by the harsh treatment and environment in the instrument. This makes direct identification of a molecule by its mass spectrum practically uncertain. For a thorough identification of a sample component, the ions corresponding to the characteristic mass of a given analyte (precursor) can be forced to fall apart in a collision chamber to fragment (product) ions, which then results in different m/z ratios that can be measured. This is known as MS/MS analysis. The m/z ratios of the fragments can be puzzled together to reveal the mass of the precursor ion. Both positive and negative ions can be measured, but if the sample components cannot be ionized, they will not show up in the mass spectrum.

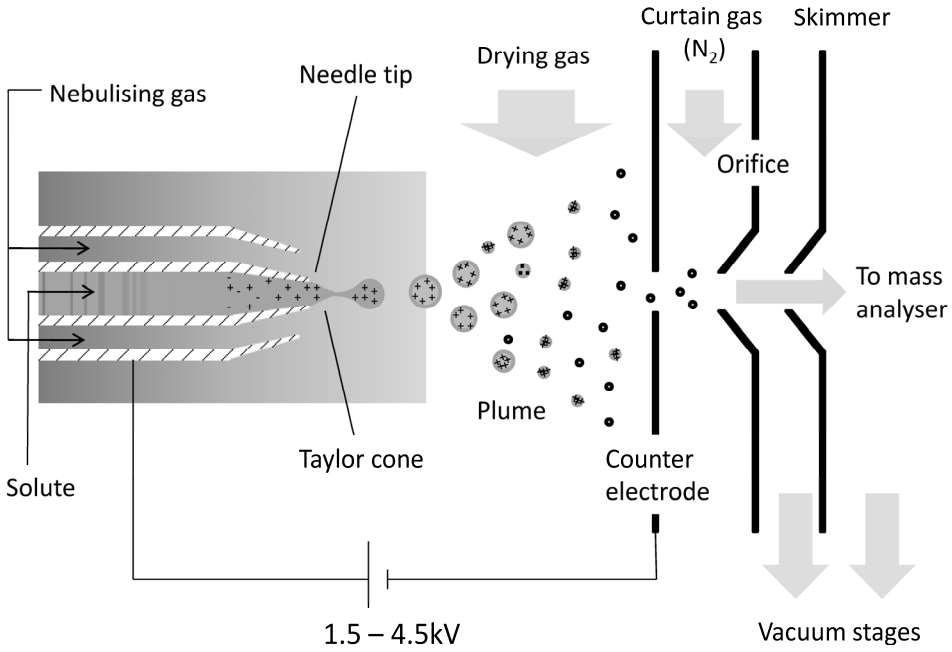

Figure 2. Schematic picture of electrospray ionisation (ESI) chamber.

In the experiments carried out in the work described in this thesis, several instruments have been employed. They all consist of an RP-HPLC coupled to a MS equipped with a quadrupole mass analyser, however, and are all hyphenated by an ESI chamber. Furthermore, the data has always been collected in full scan mode (see below).

2.3.1. Acquiring data

The instrument can be set so that only the signal from a few selected m/z ratios are registered from the samples, but when the components in a sample are unknown, the instrument is often set to scan a range of m/z ratios at a selected time interval. This yields a two-way data set, where the intensity of each point in the mass range is available at each time point. In other words, there is one chromatogram for each m/z ratio and one mass spectrum every time the signal was measured. A typical data set obtained from a quadrupole instrument, scanning in the mass range of 100 – 1000 m/z, with 0.5 amu resolution, sampled at 2.5 Hz for 30 minutes results in approximately 8 million data points. The vast amount of data is difficult to visualize and get a grasp on. The matrix of intensities depicted in **Fig. 3** contains less than 20.000 data points or approximately 0.2 % of the aforementioned example. To obtain a brief overview of the data set, the total ion chromatogram (TIC) where all chromatograms have been added, or the base peak chromatogram (BPC) where only the maximum signal in each time point is visualized, often yields the main features of the data set.

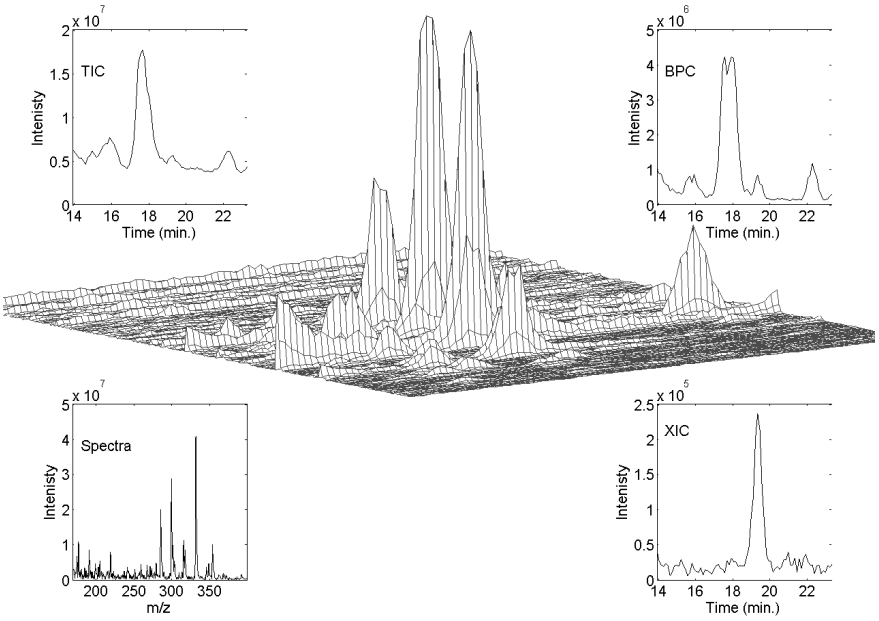

Figure 3. A part of a data set and some common visualisation techniques (TIC, BPC, XIC, mass spectrum)

It is possible that low intensity components become undetected in these condensed representations. An extracted ion chromatogram (XIC) yields the chromatogram for a selected m/z ratio, but does not show any information about the rest of the data set. The intensity and area of the peaks in the chromatograms are proportional to the concentration of the components in the sample. The spectrum shows which m/z the current chromatographic peak consist of, most often the spectra are shown for the peak apexes, or summed over the entire peaks.

3. RAW DATA PROPERTIES

The properties of the raw data are important for the degree of success of data processing, as some assumptions are made that do not always coincide with reality. The aim is to find a decent working model, and no method exists that is capable of adapting to all possible data set variations obtained by LC-MS. Regardless, the instrument should if at all possible be optimized for the current sample and the quality of the raw data should never be neglected due to a conviction that everything will be solved by data processing. The results are highly dependent on the quality of the raw data.

Some unwanted disturbances of the analytical signal are, however, always present that cannot be compensated for by tuning of the instrument. Mathematically, a LC-MS data set, **D**, can be seen as a matrix of the signals corresponding to the analytes in the sample, **A**, blurred by the additive chemical, **B**, and instrumental, **E**, noise as shown in **Eq. 1**.

$$\mathbf{D} = \mathbf{A} + \mathbf{B} + \mathbf{E} \quad (1)$$

Chemical noise arises from variations in the system or in the ambient room that are not accounted for, such as temperature, pressure, humidity, column bleeding or late-eluting compounds from prior injections. The mobile phase can also contain constituents which are continuously detected throughout analysis. This type of noise often contributes to the low frequency noise, often referred to as the baseline. Instrumental noise can be present in various forms and can arise from several sources such as pulsations of the pumping system, from the processing of the signal and by different transducers, through random fluctuations of the electric current, and from conductors that pick up and convert electromagnetic radiation into electric signals, to name a few examples. This type of noise contributes mainly to the noise of higher frequencies [4].

3.1. Ideal and non-ideal data sets

Data sets can be regarded as ideal when they have certain features that are easily and readily recognizable with the naked eye. The noise should preferably be Gaussian white (i.e. the values are independent and of normal distribution) and the baselines flat or only slowly varying. Analytical peaks should preferably be Gaussian shaped, without fronting or tailing attributes (i.e. symmetrical). The peaks should further be sampled in a way that describes the essence of the peak such that peak heights and areas are accurate, can be differentiated from the uninformative signals, and are below the maximum limit of the detector (i.e. within the dynamic range). The number of sampling points required depends on noise and peak shape. While three points are actually enough to describe an ideal Gaussian peak (one for its height and two for its width), the sampling of the instrument is commonly tuned so that 10-20 points are obtained. Theoretically, the sampling frequency should be twice the frequency of the highest frequency of the signal of interest according to the Nyquist sampling theorem [4]. Increasing the number of points augments the chance of measuring the true peak apex and area when peaks are deviating from the ideal properties, but can also increase the data sets to unmanageable sizes. A greater number of sampling points also have benefits from a signal processing point of view; the best sampled peak in general is sampled so that it contains frequencies between the noise and the baseline frequencies, and a greater number of sampling points may increase the gap between these. The new types of ultra-pressure LC-systems must manage to sample the signals at a higher rate to fulfil these requirements for achieving a reasonable accurate peak shape with a decent discriminating power. The mass spectrometer must then keep up with the higher sampling rate to maintain its usefulness as a good detector.

Another sought-after property of an ideal data set, which is of utmost importance if the components are to be quantified, is that the peaks are decently separated from each other. The resolution, R_s, is a measure of how overlapped two chromatographic peaks are and is dependent on the relative retention, t_R, and width (at base), w_b, according to **Eq. 2**.

$$R_S = \frac{2(t_{R1} - t_{R2})}{(w_{b1} + w_{b2})} \tag{2}$$

For peaks sampled in the same signal, critical pairs with an R_s >1.5 are defined as baseline separated (less than 1% overlap if equally sized) and thus become easily

differentiable and quantifiable. Partially overlapped peaks ($R_S < 1.5$) can become difficult to detect and quantify, while totally co-eluted peaks (Rs = 0) cannot be differentiated. **Fig. 4(b)** shows two examples of peak pairs with different R_s values.

In LC-MS data sets, however, totally coeluted components can be distinguished by their spectra if they exhibit different m/z ratios. Each component can give rise to several peaks though, and the information of the belonging is often limited so that an additional analytical run with different chromatographic parameters is often required to discriminate the components.

A common feature of experimental LC-MS chromatographic peaks is that they differ in width. The width of a peak can be defined in many ways, but most methods correspond to measuring the width at a certain height of the peak, such as width at half height or the width at four standard deviations ($w_{1/2}$ and w_b in **Fig. 4(a)** respectively). Since the dwell time is longer for sample components with high affinity to the column, their band broadening will generally be more pronounced compared to a component that elutes early. Lower intensity peaks also tend to obtain a somewhat smaller width than the higher intensity ones at the same elution time. This feature can influence the processing of the data sets if static peak widths are assumed. Gradient data sets, however, obtain more or less the same peak width throughout the chromatogram.

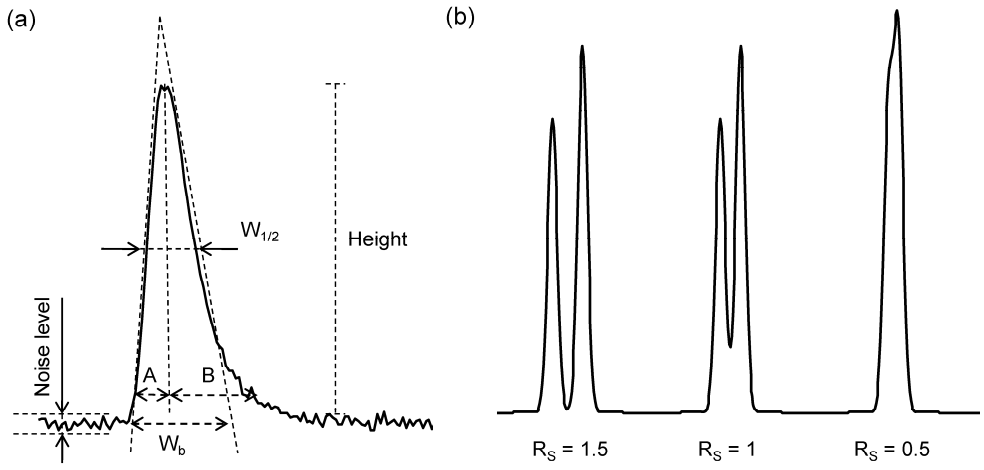

Figure 4. Some common peak properties (a) and three examples of the resolution of two adjacent peaks (b).

Another common undesirable feature is that peaks are more or less asymmetric to some extent. The most common feature is tailing, which means that some of the constituents in a component band are lagging behind in the column; this generates a peak with a normally shaped front part, but with an elongated tail. The opposite can also take place - an elongated front with a normal tail, but this is less common. Tailing can be an effect of partial clogging, an extra void or a contamination present in the column, a stronger dilution solvent compared to the mobile phase, extra column volume (unnecessary long tubing before and after the column) or by overloading the sample. Moreover, nitrogen groups in the sample can interact with uncovered silanole groups in the column packing material. Peak asymmetry is often measured by dividing the width at a certain height to the right of the peak apex, *B*, with the width to the left, *A,* as shown in **Fig 4(a)**.

Another common attribute of data sets that can severely affect the processing of data is the presence of noise deviating from being independent and/or Gaussian distributed. Measuring the high frequency noise level as the standard deviation is difficult since the signals in a LC-MS data set often also contain low frequencies (i.e. the baseline). For an accurate noise level estimation, the baseline should be levelled out before the standard deviation of the noise is measured. The peaks in a data set should reach at least 3 times above the standard deviation of the noise to statistically ensure their presence.

The properties of an ideal data set provide the basis for some of the assumptions made for many of the reported processing methods.

All the experimental data sets acquired in this thesis showed more or less severe deviations from sought-after ideal behaviour, which realistically is often the case. Data set variations and combinations are almost endless and an optimal data processing method should be able to cope with this. Synthetic data sets, with controlled deviations from the ideal case, can give complementary insight of how the methods perform insome of the non-ideal cases.

4. SIGNAL PROCESSING

In addition to controlling the instrument and acquiring data, computers can be used to aid the analyst in the extraction of relevant information from the data sets and during method development by increasing the signal to noise (S/N) ratio, detecting and controlling the purity of a peak or tracking components when chromatographic conditions have been changed.

In all LC-MS data sets some degree of noise is always present. If the noise is abundant, it can be difficult to detect the analytical signal. The S/N ratio is a

measure of the extent of signal corruption by noise. The S/N ratio can thus be improved by increasing the signal, decreasing the noise or both. Common signal processing methods, sometimes developed for completely different scientific fields, have successfully been applied to LC-MS data. The role of signal processing in analytical chemistry can be traced to the development of instruments capable of measuring and storing a continuous signal, the analog-to-digital converter (ADC), and the development of efficient digital signal processing methods [5].

4.1. Digital filtering

Digital filtering is a form of signal processing by discrete methods that perform mathematical operations to manipulate the sampled signal. It is one of the most widely used methods for signal processing in analytical chemistry [5]. Several types of digital filters exist, but perhaps the simplest and most commonly used are the non-recursive filters, in which the discrete first-order raw data signal, y, is convoluted with the filter coefficients, c, according to **Eq. 3;** the output signal, y', is not used as input during progress. The output point becomes an estimate of the current point to be filtered in the unprocessed data and its neighbouring points. The shape of the output signal is often affected by the characteristics of the filter coefficients.

$$y'_j = \sum_{i=-m}^{i=m} c_i y_{j+i} \qquad (3)$$

The filter can be applied to operate in the spectral domain, in the chromatographic time domain, or in both simultaneously. Since the peaks of interest are generally wider in the time direction of the LC-MS data set, the analytical signals are normally more easily discriminated from the noise in the chromatographic time domain.

4.1.1 Filter coefficients

A digital filter acts on the current point to be filtered together with an arbitrary number of the neighbouring points by summing fractions of their original values. The filter coefficients of a digital filter can be seen as weights determining how much influence each point in the window should have on the output signal. The product is calculated between each data point and corresponding filter coefficient

and the sum of the products becomes the new filtered point. With the standard procedure, the window is moved to the next point and all points included in the previous calculation are still present, except for the last one which is discarded and replaced with the next unprocessed point in the series. The process is schematized in **Fig. 5**. Most commonly in LC-MS applications, the values of the filter coefficients are static and symmetrical with the highest weight at the centre and the window width is fixed throughout the filtering process. The coefficients are often normalised to unit sum to obtain a decreased noise level and constant signal height, or to unit length to obtain a constant noise level and an increased signal. The actual increase in S/N is independent on type of normalisation. Some common filter coefficient functions are depicted in **Fig. 6(a)**.

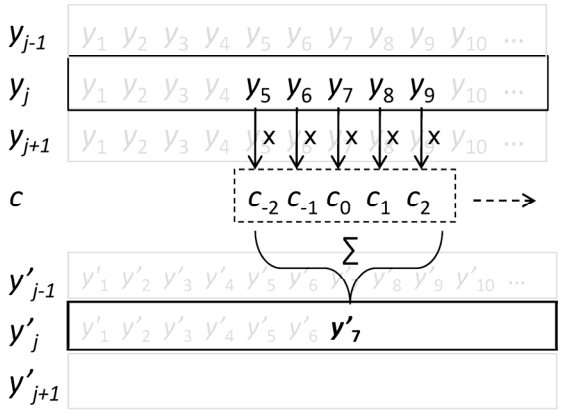

Figure 5. Workflow of a non recursive digital filter, where m = 2, currently working on the kth chromatogram and with current value of j = 7.

The resulting signal will also in most cases benefit from a smoothing effect, where the high frequency noise in the peak will be embedded in the new smooth filtered version. The filter will, however, also influence specific frequencies of an noise-free signal [6]. As a consequence, peak shapes often becomes slightly different. If noise is available in all frequency bands, filtering without peak distortion is impossible to achieve [7]. This side-effect can have an influence if the signal is intended for use with multivariate calibration, for example [6]. Peak distortion also includes changed peak widths, which influence the chromatographic resolution. Often the coefficients can be set to allow for greater

S/N improvement with the drawback of a decreased resolution, or a constant or even improved resolution at the cost of a lesser S/N improvement.

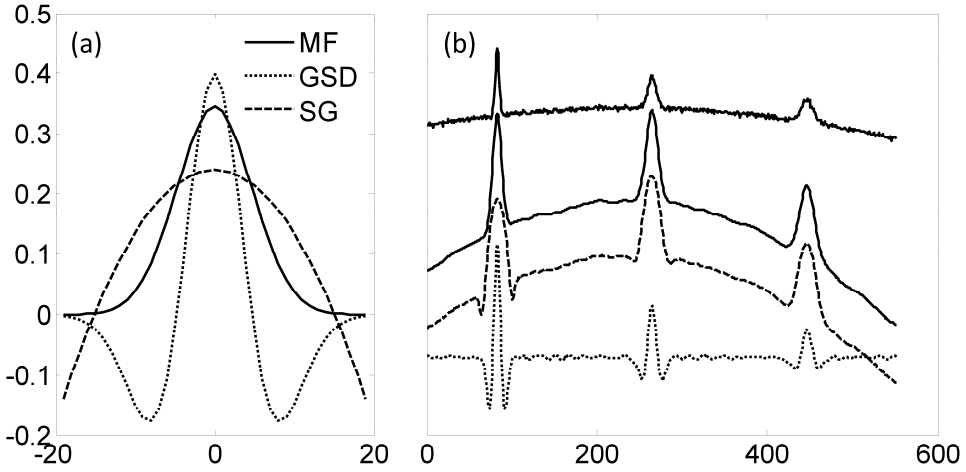

Figure 6. (a) Some common filter coefficients, matched filtration (MF), Gaussian second derivative (GSD) and 2nd order Savitzky-Golay (SG). All are normalised to unit length and consist of 41 points, σ = 4.75 for MF and GSD. (b) A simulated chromatogram (top) and the effect of applying the specified filter coefficients. The widths of the three peaks in the simulated chromatogram are close to optimal regarding S/N enhancement for the coefficients in (a) for GSD = left peak, MF = middle and SG = right. The area of the peaks in the simulated chromatogram is equal. The filtered versions are separated for clearer visualisation.

4.1.2 Common types of filter coefficients

The theoretically best result in terms of S/N improvement are obtained when the function of the filter coefficients equals the analytical signal as much as possible and, conversely, does not match with the noise or background signals. In the case of LC-MS data, a Gaussian function of the filter coefficients would then increase the S/N level the most, if the data are ideal with white noise and perfectly shaped peaks [8-10]. These filters are commonly referred to as matched filters (MF).

A moving average filter, on the other hand, has flat static coefficients so the output becomes simply the average of the neighbouring points [11]. This results in a smoothing effect, but does not work well to retain the shape of the analytical signals in LC-MS data [9,12]. Filtering with several window widths simultaneously can improve the results [13].

The Gaussian second derivative filter (GSD) is the negative of the second derivative of the Gaussian function [14]. This set of filter coefficients matches with a Gaussian peak to some extent, and at the same time has edges below zero to result in the sum of the coefficients equalling zero. This way, a total reduction of the baseline to the zero line is achieved, which can be a very nice feature when some or many of the baselines are highly fluctuating and therefore interfere during data set overviews such as in TICs or BPCs.

In Savitzky – Golay (SG) filtering [15,16], the filtered data point is the result of applying a least squares polynomial of selected order to a odd numbered window around the data point to be filtered. Luckily, a set of filter coefficients exists that can be used in the same manner as the other digital filters as in **Eq. 1**, regardless of filter window width, or order of polynomial. Trials have been reported where the optimal degree of the polynomial is adapted to the signal [17,18] or optimal window size [19], for even numbered coefficient windows [20] and in combination with a median filter [21].

An example of applying the MF, GSD and 2nd order SG filter is found in **Fig. 6(b)**.

4.1.3 Filtering ideal and non-ideal LC-MS data sets

Most filters assume more or less ideal data sets with Gaussian peaks, a slowly varying baseline and white noise. When assumptions about ideal data sets do not coincide with reality, it often results in deteriorated filter performance. The theoretical maximum S/N improvement can be calculated for ideal chromatographic peaks according to **Eq. 4**, where n is the number of sampling points describing the chromatographic peak and p is a factor proportional to the correlation between the filter coefficients and the chromatographic peaks.

$$\frac{S}{N}improvement = p\sqrt{n} \qquad (4)$$

If MF coefficients are used with optimum width and the data set is ideal, p will receive its maximum value of 0.67 when filtering a chromatographic peak [14]. This means that to obtain an actual improvement in S/N, the number of sampled data points of the chromatographic peak only has to exceed two, which is normally the case. The other types of filter coefficients will all yield lower values of p at optimal settings. If the peaks deviate from the ideal, or different settings of the filter coefficients are used that do not match optimally, the value of p will decrease.

If the width of the matched filter coefficients is smaller or wider than the chromatographic peak, the resulting S/N improvement will be reduced as a consequence. An actual improvement is, however, obtained for the MF filter even though the filter coefficient width differs by as much as 20 – 700 % for an ideal data set with a peak sampled with 8 data points; a wider peak tolerates even larger differences. Setting the coefficient width too narrow or too wide can, however, also enhance the noise and baseline respectively.

Different filters will also influence different frequencies of the data. For example, a 20-point MF filter will increase the S/N ratio for any peak with a width above five points, assuming ideal data. This is also true for the 20-point GSD filter, but then the effect decreases and peaks wider than 35 points will not be enhanced by the filter. The GSD acts as a band-pass filter, whereas the Gaussian matched filter acts as a low-pass filter [14].

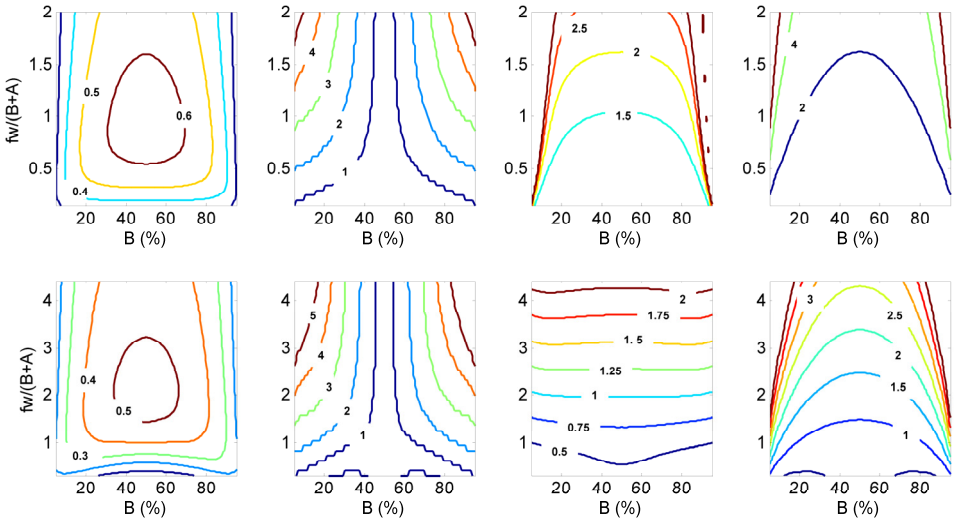

Figure 7. The theoretical values of the enhancement factor *p* (left column), output peak apex displacement (middle-left), output peak width divided by original peak width at 10% of peak height (middle-right) and at 50% of peak height (right) for different peak mismatches (y-axis) and the peak asymmetry factor at 10% height (x-axis) after matched filtration (top) and GSD filtration (bottom) of a peak where B+A = 21.

If the optimal S/N improvement should be obtained throughout the data set, the width of the filter coefficients of a matching filter also have to be adopted accordingly to the peak widths. Since the peak widths generally increase linearly with elution time for isocratic data sets [22], a couple of typical peak widths can be

measured and the others can be predicted by a linear model. For optimal coefficients regarding S/N though, the output peak width becomes approximately 40 % wider for the MF filter, which corresponds to a loss in chromatographic resolution.

Asymmetric peaks will also influence the filter performance. The perhaps most devastating effect is that the peak apex positions can change after filtering. If all peaks tail to the same extent, the effect is often surmountable but can influence more downstream when closely eluting peaks are assigned to their respective components. The more the peaks tail, the less the shape of the filter coefficients coincides with the peak, and p and thus the S/N improvement are reduced. The resulting filtered peak is often less skewed than the original peak if the filter coefficients are symmetric around its centre point. The difference in width and asymmetry before and after filtering depends on which definition is used for the peak width.

In **Fig. 7,** the value of p, the peak apex displacement, and peak width after filtering are shown. The results are exemplified for a noise free artificial peak with a constant width at 10% height, but with a different peak asymmetry factor. The effect on the different variables is also shown simultaneously when applying different filter coefficient widths.

The properties of the noise also affect the filter performance. If the noise or baseline has frequencies within the frequency band of the filter, their intensity will be enhanced as well. In some circumstances, this can result in reduced improvement or even a reduction in S/N level after filtering. This and other situations where the performance of these types of filters often is deteriorated are exemplified in **Fig. 8.**

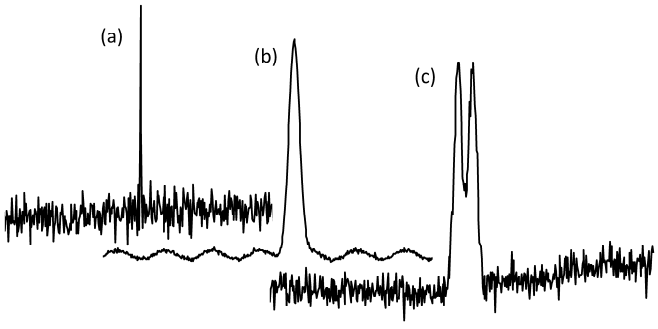

Figure 8. Some examples where a symmetrical non-recursive filter often fail to acceptably enhance the S/N ratio even though optimised for the current peak width: (a) narrow peaks, (b) baseline or noise with frequencies within the frequency band of the filter, (c) adjacent peaks.

4.2 Other common LC-MS signal processing methods

The component detection algorithm (CODA) is a popular and fast data set reduction technique, where the quality of each chromatogram acquired is estimated by comparing the original data with its smoothed and mean-centred version [23]. Chromatograms are completely discarded if their quality is below a user-decided threshold. This removes uninformative signals and enhances mainly the TIC and BPC representations. Variants and improvements of the CODA algorithm have been reported [24,25].

Moreover, filtering of LC-MS data has been reported after transforming the data into another domain. This subject requires more elaboration than the framework of this thesis allows. The insight that a signal can be seen as a combination of sine waves enabled the possibility to convert the signal into the frequency domain [12]. This generates data in a different form that can be manipulated in a different manner, but essentially contains the same information. This has resulted in a huge amount of applications in many scientific fields where signal filtering is one. A drawback of such transformation is that the time information is lost (i.e. manipulations made in the frequency domain influence the entire signal when transforming back to the time domain). Therefore, alternative transforms that represent signals in the trade-off between optimal time resolution and optimal frequency resolution exist. Wavelet transformation is such an example that has increased in popularity among analytical chemists during the last two decades [26], and has reported applications beyond filtering for chromatography data [26-34].

Other reported methods for increasing S/N include differentiation [35,36] and multiplication of neighbouring spectra [37,38]. These methods suffered, however, from severe peak distortion or required somewhat ideal data. Noise reduction can also be performed by only retaining the most significant principal components in principal component analysis (PCA), the basic principles of which will be briefly explained in a chapter below [39].

In *paper I*, various methods (CODA, GSD, MF and SG) for improving the S/N ratio were applied to investigate their effects on experimental and simulated LC-MS data in terms of improvements in the TIC, BPC and XIC representations. It was found that the enhancement with the experimental data was meagre and it was for this reason that the data sets deviated from the ideal in some aspects (e.g. **Fig. 8**). Even though the analytical signal was enhanced, so was also the noise. Thus the S/N ratio improvement was sparse or absent for the XICs. The noise contained

frequencies that were enhanced by the filter, even at optimal filter settings. The actual optimal coefficient width also deviated from the theoretical.

GSD appeared to have some interesting features though such as a reduced baseline, which also improved visualisation and interpretability of the mass spectra. Furthermore, the S/N in the TICs and BPCs were improved and the filtered chromatographic peaks received a width close to the original when the filter was applied at optimal settings in terms of S/N improvement. It was believed that these features could be utilized in the next step for reaching the goal of the project, namely to detect the peaks and discard the uninformative parts of the acquired signals.

5. PEAK DETECTION

Peak detection includes methods to localize the informative peaks in the data. The detection of peaks can be achieved in the chromatographic time domain or in the spectral domain. Theoretically, it is preferable to work in the time domain for LC-MS data since it is easier to distinguish the informative signal from the noise in the chromatograms [28,40], and since there is a lower risk that the peaks are distorted in the spectral domain by the mathematical operations applied [41]. In the spectral domain the peaks generally have the same frequency as the noise and can only be differentiated by a higher intensity. A chromatogram containing only a high baseline column could be mistaken for an informative peak in such spectra. Sometimes it is desirable to detect peaks online as the elution progress, and then peak detection in the spectral domain is the natural approach. Other methods claim that peaks are best extracted by utilizing both domains simultaneously.

Common peak detection algorithms working in the chromatographic domain are often capable of extracting properties other than retention time, such as peak width, area and symmetry. Often the attributes of the peaks are stored in a peak list after peak detection, which can be used for further processing. Alternatively, the gained information can be used to reconstruct the data without the noise or baseline contribution. Due to the vast number of data, manual peak detection can, however, be very time inefficient, tiresome and thus subjective and prone to error. The optimal peak detection algorithm would collect data solely from **A** in **Eq. 1**, and no residuals from **A** should exist. This utopic scheme is difficult to obtain in reality, however. Some risks associated with peak detection include the discarding of informative parts or the presence of false positives.

One of the earliest, simplest and most commonly applied methods for peak detection is to collect data above an arbitrary intensity threshold. It is only plausible however if the noise has essentially constant amplitude and the baseline is flat and of equal height throughout the data set. Since this is highly unlikely, the problem can be solved by changing the threshold as the noise amplitude and baseline are varying. If the baseline can be settled and removed from the data, the noise estimations become easier, since the standard deviation of an arbitrary interval of the noise is dependent on the slope of the baseline. The resulting peak height and area will better represent the actual content of the corresponding sample component. Some robust processing of the data is needed, though. If properly implemented, this classical method functions satisfactorily and is easy to understand. The typical peak detection steps based on this strategy are shown in **Fig. 9**.

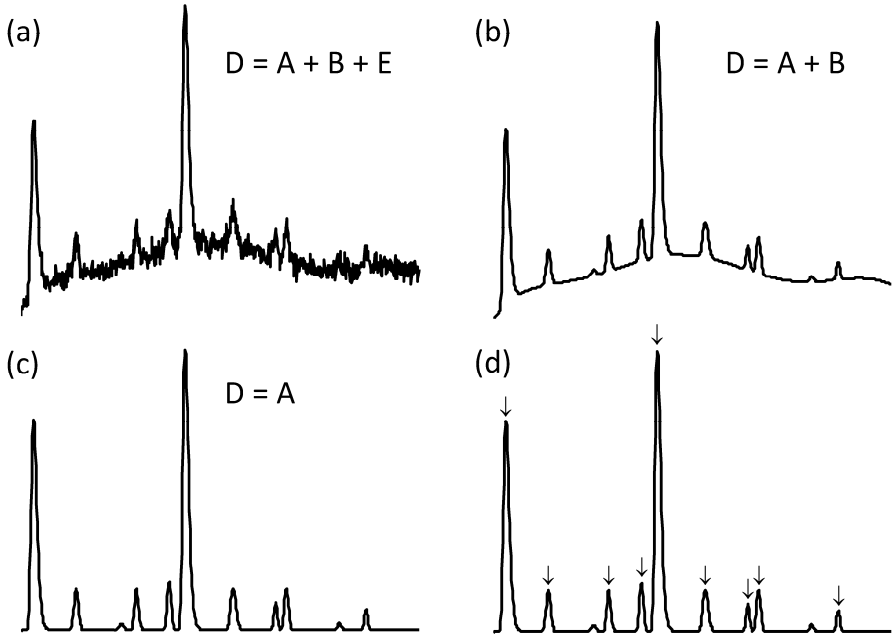

Figure 9. Typical workflow of classical peak detection. (a) The original chromatogram, (b) reduced noise, (c) reduced noise and baseline, (d) Peaks detected above a threshold based on the noise.

5.1 Baseline reduction

If the original peak shapes are important to preserve, the baseline can be withdrawn from the rest of the data by first applying an estimate to the imaginary, slowly varying signal which operates as a fundament for the informative signals (**Fig. 9a-b**) and then subtract it. It is important however that the implemented method is unaffected by the analytical peaks but also capable of adapting to abrupt changes when needed. Subtracting the running minimum [42] or median from the signal [43] or adopting a slowly varying function of arbitrary degree in the least squares sense to each chromatogram may function well when the shape of the baseline is anticipated, but often generates unacceptably large residuals when deviations occur. Results can nonetheless be improved by iterative approaches [44].

When a baseline varies from linear to complex within the same chromatogram, piecewise polynomials, also called splines, can be used with convincing results [45-47]. These consist often of quadratic or cubic polynomials adapted to portions of the chromatograms and joined together. By adapting to portions, the variability is reduced in comparison to the complete signal. By assuring that each polynomial has the same slope and curvature at each junction (i.e. the same first and second derivative), the resulting baseline estimation becomes thus adaptive but also smooth. Each piecewise polynomial is influenced by an arbitrary number of data points whereas the rest is interpolated. The key to successfully applying the spline is to avoid the data points describing the peaks. This makes automation more difficult however, since peaks can exist in practically any position in the data. One method for overcoming this problem is the use of a rational spline that has additional weights associated to each of the influencing points. The location of the peaks in the data can be roughly estimated and the influence from these points can be totally reduced. By doing so, all data points can be used to influence the spline at the positions where peaks are not present. In this way, the adaptation is nearly optimal at these positions. Unpublished results using rational Bezier splines showed generally excellent adaptation to highly fluctuating baselines, but showed some apparent glitches for very wide peaks in some circumstances. An example of a successfully adopted baseline with a rational Bezier spline can be seen in **Fig. 10**.

If a preserved peak shape is of lesser importance, the baseline can be removed by utilizing the derivatives of the signal. The first and higher order derivative substantially increases the noise and distorts the peaks, however. Smoothing prior differentiation can reduce or even increase the S/N ratio and using the negative second derivative generates peak-like shapes. The derivatives of a signal can also

be utilised to find the apex and start and end points of a peak [48]. Wavelet transformation with a symmetrical wavelet, such as the Mexican hat, automatically reduces the baseline in a manner similar to the GSD filter (**Fig. 10**), or any other symmetrical filter with the sum of the filter coefficients equal to zero.

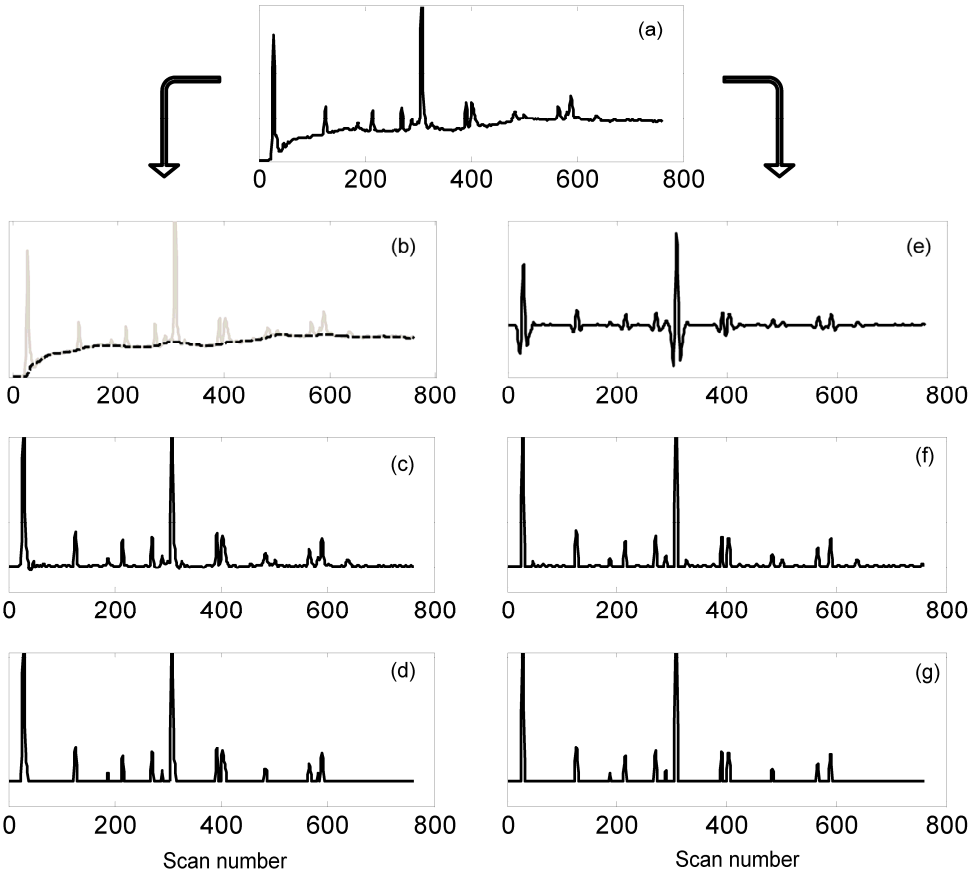

Figure 10. Left column: Peak detection with rational Bezier spline. Right column: Peak detection with GSD filtering. (a) The original chromatogram, (b) The baseline (dashed) adapted with the use of splines, (c) baseline removal, (d) detected peaks > threshold. (e) GSD filtered representation, (f) GSD signal > 0, (g) GSD signal > threshold.

5.2 Estimating the noise level

With a reduced baseline, the noise level in a chromatogram can commonly be estimated by measuring the standard deviation of the noise, but other definitions of the noise level exist as well. The amplitude of the noise can, however, change during a LC-MS run in some of the chromatograms. To obtain decent statistical precision, the calculation of the standard deviation requires a finite number of data points. The noise level should thus be estimated in the region of the peaks, commonly covering 20 times the peak width (w_b or $w_{1/2}$) [49]. The noise regions are, however, difficult to establish without knowing the location of the peaks and vice versa. To automatically estimate the noise levels, a rough estimate can first be used on a baseline levelled chromatogram such as the median absolute deviation (MAD) which is the median of the absolute values of the deviations from the median of the data [47]. The MAD value is more resilient to outliers in the data compared to the standard deviation and in chromatography; the peaks can be regarded as outliers compared to the noise. The regions of noise must be larger compared to the region of signals for this to be applicable, though. For normal distributions, the standard deviation can be estimated from the MAD value through multiplication by 1.483. For an improved accuracy, the signals above three times the obtained product can then be discarded before recalculating the noise level by the standard deviation of the remaining signal locally in the neighbourhood of the peaks.

Other noise definitions have been reported such as average random deviation divided by the square root of the signal intensity [50].

5.3 Extracting peaks

Once the noise level has been established, peaks can be extracted. As a general rule, the limit of detection in analytical chemistry is based on signals three times above the noise, that is a $S/N \geq 3$, whereas the limit of quantification are often set to $S/N \geq 10$ [51]. These limits are derived from the possibility of obtaining type I and type II errors from basic statistics. Using a higher S/N limit decreases the risk of obtaining false peaks (type I error), whereas lower S/N thresholds decrease the risk of missing minor analytical peaks (type II error).

5.4 Other peak detection methods

Other automatic methods for finding peaks in LC-UV, MS, GC-MS and LC-MS data have been reported in the literature, with widely different approaches, by detecting peaks in the chromatographic domain [40,41,52], the spectral domain [7,28,29,42,53-60], or both [46,61-63]. Some of the methods utilize morphology [54,63], other extract peaks from the GSD filtered signal [40], or use specific statistical methods [60], whereas some of the more recently developed methods include transformation of the data in to the wavelet domain prior to peak detection mainly to remove noise and/or background in the chromatographic domain [28,46,52,53] or spectral domain [29,42,55,59,61]. Some utilize the information from typical isotopic patterns [46,54,61]. A comparison of some publicly available programs for peak detection and their functionality can be found in [43].

In *paper II*, the GSD filter was applied to the data as a first step to facilitate peak detection. This resulted in some interesting benefits. By using the negative second derivative of a Gaussian function, the resulting signal is smoothed, obtains a higher S/N and a reduced baseline. The filter functions as a band-pass filter so both high and low frequency signals are damped. The GSD filter is less computationally intense compared to using rational Bezier splines for baseline removal for example, a method that was also tested. Moreover, the reduced baseline facilitates decent noise measurements and with filter coefficients normalised to unit length, the noise levels are retained for ideal data sets. In addition, the width at base is not changed considerably at optimal S/N settings after filtering. The noise could thus be estimated in the manner outlined in **section 5.2** and the peaks could be harvested at some arbitrary threshold. At optimal S/N settings, the GSD filtered signal has the drawback of an obtained decrease in resolution due to the fact that the negative parts of the coefficients is not inside the boarders of the peak to be filtered. When it comes to GSD and many other digital filters, there is a trade-off between optimal S/N and resolution, and the filter coefficients can be altered so that the resolution is retained or even increased at the cost of a decreased S/N enhancement. The shape of the resulting peak is not precisely Gaussian, but has the peak apex at the correct position, though asymmetry of the original peak can cause deviations. Non-ideal noise containing frequencies within the band-pass properties of the filter will be increased and the S/N improvement may be reduced or even lost. This occurs more often when the chromatographic peaks are deficiently sampled however, as was found in *paper I*.

For optimal performance, the algorithm requires the typical peak width. The peak widths in isocratic LC-MS data theoretically increase linearly with time [22]. This means that two peak widths are sufficient to model the linear relation. Since deviations commonly exist in experimental data, more peak widths should therefore be measured to obtain a statistically sound model. A method for automatic determination of the typical peak widths in LC-MS data is presented in *paper II*, where initially the CODA algorithm is implemented to extract the chromatograms of highest quality. The peaks in the selected chromatograms had their widths estimated by using a relation between the height of the second derivative of a Gaussian peak and its width [64]. This was utilized by applying several GSD filters to the signal (this procedure is somewhat similar to wavelet transformation with the Mexican hat wavelet) and using the relative change in peak height as an estimation of the width.

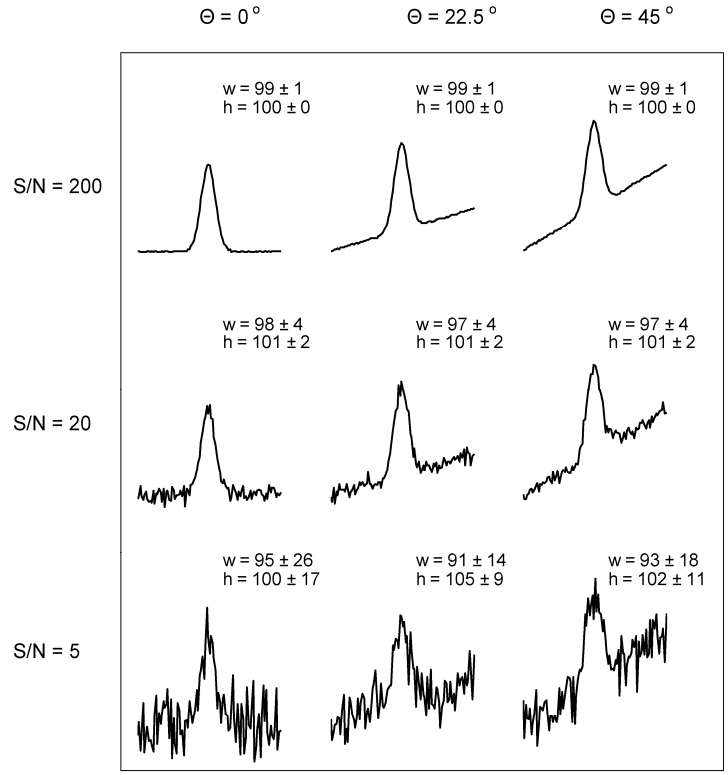

Figure 11. A simulated chromatographic peak with constant width and height at different slopes and noise levels. The letter w and h represents the ability to estimate the true width and height respectively in % of the true value. The average and confidence intervals (95%) are estimated after simulating the noise 100 times.

This method is also capable of estimating the height of the peaks, even when a baseline is present. A typical result is illustrated in **Fig. 11**, where the width and height for a synthetic chromatographic peak have been estimated by the method at different applied noise levels and slopes of the baseline.

The output signal will, however, theoretically obtain a higher increase in S/N for the wider peaks due to its √(n) dependence according to **Eq. 2**. These differences can nonetheless be accounted for by normalising the obtained noise and baseline free peaks accordingly. On the other hand, the band pass properties can cause different enhancement of the same type of non-ideal noise throughout the chromatogram as the filter progresses in time.

The resulting peak detection algorithm can thus fully automatically detect and extract peaks in LC-MS data sets with varying peak widths, noise level and baselines. A part of a peak detected and reconstructed data set can be viewed in **Fig. 12**, which was performed on the same data set as depicted in **Fig 3.** The reconstructed noise and baseline free data sets with smooth peaks were believed to increase the accuracy in the next step of the main plan, namely bunching the peaks to its corresponding components.

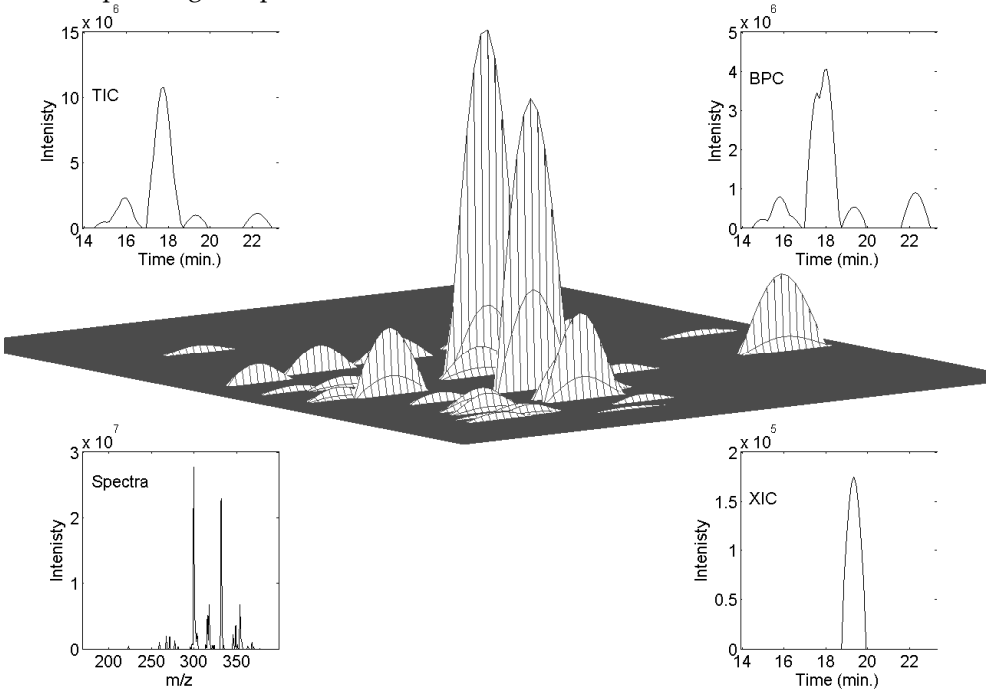

Figure 12. Part of a reconstructed peak detected LC-MS data set (c.f. **Fig. 3**).

6. PEAK CLASSIFICATION

To identify a specific component in a sample, its corresponding spectrum can be used as a more or less unique fingerprint. To be able to assign each component in a sample a specific spectrum, a proper method should be used that manages to determine which peak belongs to what component, even when the components are closely related with similar mass spectra and chromatographic performance. This is an important step which can be rather difficult to carry out during certain circumstances, especially automatically.

6.1 Peak purity

The underlying problem formulation for bunching analytical signals into respective components is closely related to a highly developed and devoted subject in the chromatographic society: how to detect and separate impure peaks. Impure components are sample components that coelute during the applied chromatographic conditions. Because of this problem, some of the defined analytes in a LC-MS data set can stay undetected whereas others can obtain exaggerated abundance and confusing characteristic spectra. Two or more actual components containing exactly the same spectra that elute at exactly the same time cannot be separated by any method in a single data set without further experiments. Further, components that are embedded in other components to some extent are difficult to discern. Mathematically, the goal is to divide the analytical signal **A** (*n*x*m*) into component chromatograms, **C** (*n*x*q*), and spectra, **S** (*m*x*q*), according to **Eq. 5**, where *T* stands for the transpose of the matrix **S**, **E** is the measurement noise, *q* is the number of components in the sample and *n* and *m* are the size of the rows and columns of the matrix **A**, respectively.

$$\mathbf{A} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad (5)$$

Methods have been developed that can be used to separate partially and severely coeluting peaks by obtaining a solution to **C** and **S** in **Eq. 5** as visualised in **Fig. 13**. Such deconvolution problems are not straightforward however, due to inherent uncertainties of the shapes of **C** and **S**. Many methods were originally developed for absorbance data where spectra are less specific compared to mass spectrometry data [65].

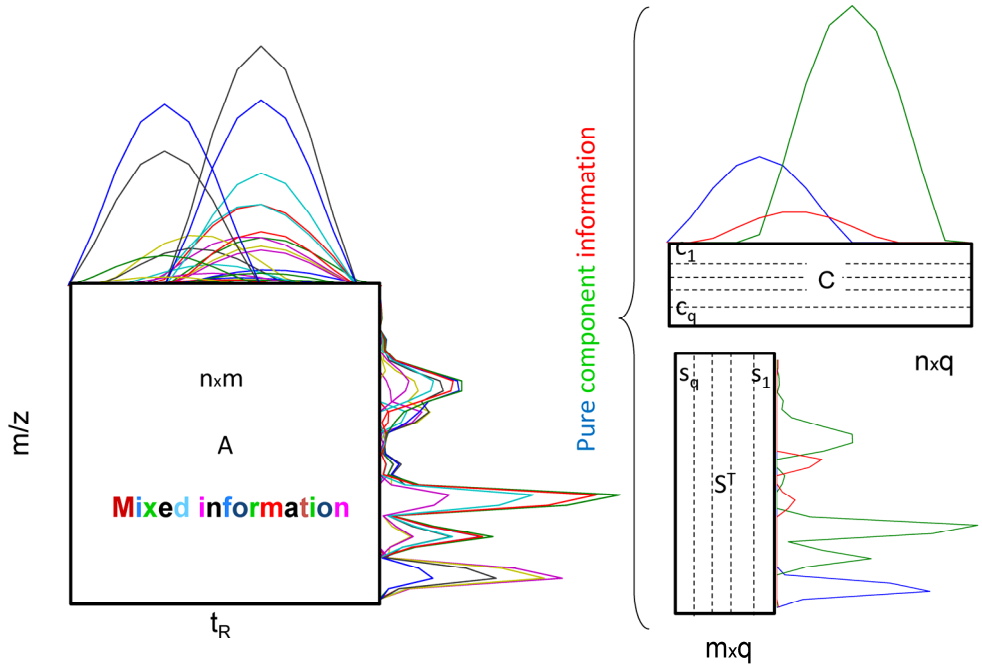

Figure 13. A data subset and graphical representations of C and S (**Eq. 5)**

Some of the most common and successful methods that can automatically deconvolve the data are the multivariate curve resolution (MCR) or self-modelling curve resolution (SMCR) techniques, which essentially require only the information from the data in **D** (or **A**) to function and are thus easier to automate. In these methods, **C** or **S** are first estimated and further improved iteratively. Multivariate curve resolution alternating least squares (MCR-ALS) [66,67] and iterative target factor analysis (ITTFA) [68,69] are two popular non-parametric approaches. In MCR-ALS, the least square estimation of **S** from the data and an initial estimation of **C** (or vice versa) are first established. Then **S** is manipulated before **C** is re- estimated by least squares. This then progresses until further iterations no longer improve the results. In ITTFA, estimated component chromatograms, or targets, in **C** are iteratively refined one at a time, until the target match satisfactory with one of the true profiles. The spectra are then calculated by least squares from the data and estimated component chromatograms in a single step. Two classes of ambiguities are associated to these methods, namely rotational and intensity ambiguities. Rotational ambiguities arise due to the fact that the obtained solution to **Eq. 3** can mathematically be an unknown linear combination

of the true solution according to **Eq. 6**, where **R** is an arbitrary transformation matrix and **C′ = CR** and **S′=R**$^{-1}$**S**^{T} are the obtained solution [70,71].

$$\mathbf{A} = \mathbf{CRR}^{-1}\mathbf{S}^{T} + \mathbf{E} = \mathbf{C'S'}^{T} + \mathbf{E} \qquad (6)$$

Due to the rotational ambiguity, the resulting component chromatograms and spectra can be mathematically sound but obtain a shape that is unlikely or even impossible in reality.

Intensity ambiguities can be explained by rewriting **Eq. 5** according to **Eq. 7**.

$$\mathbf{A} = \sum_{i=1}^{n}\left(\frac{1}{k_i}c_i\right)\left(k_i s_i^{T}\right) \qquad (7)$$

Here, the correct shape of C and S can be obtained, but with k_i smaller or higher intensities of C or S, depending on whether k_i is above or below one. The ambiguities can be decreased by using constraints to steer the process between iterations into chemical meaningful solutions [67,70,72]. Normalising the spectra between iterations suppresses the intensity ambiguity, whereas rotational ambiguities are more difficult to control even though several types of constraints have been developed; this is still a hot topic in this research area. The extent of ambiguity depends on which level the components overlap at in the data and can therefore vary greatly within a data set.

Other curve resolution methods exist that solve the deconvolution problem non-iteratively [70,73,74], or via parametric models [75]. A variety of methodologies for MCR are outlined in [70,76] and a paper where several MCR methods are compared can be found in [77].

Curve resolution can be extended to be simultaneously performed on several, augmented, data sets [72,78,79]. This is possible as long as either the chromatograms or spectra are equal between runs. Overlapping peaks can then be resolved if they have selective regions in at least one of the individual data sets and intensity ambiguity are unessential since the relative intensities are preserved. Data sets can even be augmented with data from different detectors [80-83], which then often require the same number of scans in time, or some truncating of one type of data to fit to the other in at least one dimension**.** Hyphened strategies could theoretically be used with the types of data sets obtained from the typical LC-MS method development. It assumes equal spectra for the same species in the different experiments, which cannot be assured due to differences in temperature, ionic strength and solvents etc. The extremely large data sets commonly obtained when

fusing these kinds of data sets becomes difficult to handle though, even for a modern computer.

6.2 The chemical rank

Many of the methods developed to solve Eq. 5 first involve the determination of the number of chemical components in the system. The number of components is commonly estimated by determining the mathematical rank of a series of subsets of the data, which preferably should contain a small number of complete sampled components, through one of several method developed for this purpose. The method used is successfully if the estimated mathematical rank equals the chemical rank of the subset. Subsets are used to reduce the impact of varying peak widths, when present (e.g. isocratic), and colinearity which occurs when the spectra or component chromatograms are equal or very similar for two components. Colinearity makes the mathematical rank decrease and thus impairs the results from the employed method. Noise and background also blur the edge between informative and uninformative eigenvalues with an increased risk that the chemical rank of the system is erroneously estimated.

6.3 Factor analysis

The mathematical rank can be determined by factor analysis, where principal component analysis (PCA) is a commonly used method. PCA rearrange the data set so that the first principal component lies in the direction where it describes the most of the variation in the data. The succeeding component is then drawn perpendicular to the previous component and then describes the most variation of the rest of the data. By doing so, the new variables become uncorrelated and this can be solved mathematically by determining the eigenvalues and eigenvectors with eigenvalue decomposition or by singular value decomposition. In the case of LC-MS data, each m/z value is considered as a variable and each scan time as an object. After PCA the new variables, called principal components (PCs), are linear combinations of the original variables with decreasing importance. The number of obtained principal components with an eigenvalue significantly above the eigenvalues of the noise components is the mathematical rank of the system. Each cluster of chromatographic peaks eluting at the same time and shape will contribute substantially to the variation of the original data and point in the same direction in the PC space. Noise contributes little to the variation of the data but is often abundant and thus generates a lot of smaller principal components. Noisy components can be removed, but the difficult aspect is to set the threshold value on the eigenvalues that correspond to the importance of the PCs.

6.4 Initial estimations

As mentioned earlier, many curve resolution methods require an initial estimation of the component chromatograms or spectra. Even though a totally randomized estimation is sufficient in some cases, studies have shown that estimations closer to the true solution generally generate better results [77,78,84-86]. Some of the methods for estimating the initial component chromatograms and spectra utilize PCA for local rank analysis and to visualise the number of components in the subset. Evolving factor analysis (EFA) uses information not available from classical factor analysis by calculating the eigenvalues and plotting their logarithms as elution progress in the data subset, by increasing the number of included spectra, both forward and backward [87,88]. The resulting map shows the appearance and disappearance of components in the data, and the number of components and their position in time can be established by assuming that the first appearing component disappear first as well. Fixed size window factor analysis (FSW-EFA) is a variant where the number of spectra is fixed, so the eigenvalues are calculated in a fixed size interval that slides to the succeeding point in the chromatogram [89,90]. This results in better sensitivity to minor components and reduces calculation speed. Heuristic evolving latent projections (HELP) uses the information from the data where selective and zero-concentration regions are present, and then strips the corresponding components from the data set before new such regions are considered. These are some of the earliest and most commonly used methods, and many other methods are based on these approaches [70]. Interpreting the loadings and scores from the PCA manually can also reveal embedded peaks. There are also methods not based on PCA such as simple to use self-modelling mixture analysis (SIMPLISMA) [91].

Reviews of the methods described above for estimating the number of chemical components, the briefly explained MCR methods and other similar methods can be found in [70,92,93].

PCA was also employed in *paper III*, to estimate the number of chemical components in the sample. The reconstructed data from the previous peak detection step was divided by selecting the most abundant peak in the data and included all peaks whose retention time lied within the full width of the selected peak. The number of components was then estimated and the component chromatograms and spectra were determined with the use of a method referred to as principal component variable grouping (PCVG) [94]. This method searches for the longest vector in the PC loading space and produces initial component chromatograms by including all peaks within a specified angle of the vector. These

preliminary component chromatograms then becomes a template of the main feature of the subset which all other involving peaks are compared to and assigned depending on similarity. The assigned components are then removed from the original data and the next most abundant peak settles the border for the next subset. **Fig. 14** demonstrates the result of using this method for the data shown in **Fig. 12.** A drawback of such a method is that adjacent peaks within the same XIC, unable to be separated correctly by the peak detection step, cannot be separated during this step as well. An implementation of MCR-ALS could theoretically handle this. On the other hand, the MS spectra of closely overlapping peaks can be totally specific and an additional multivariate method is therefore redundant. Furthermore, ALS can cause unresolved spectra that become difficult to compare between data sets [75]**.**

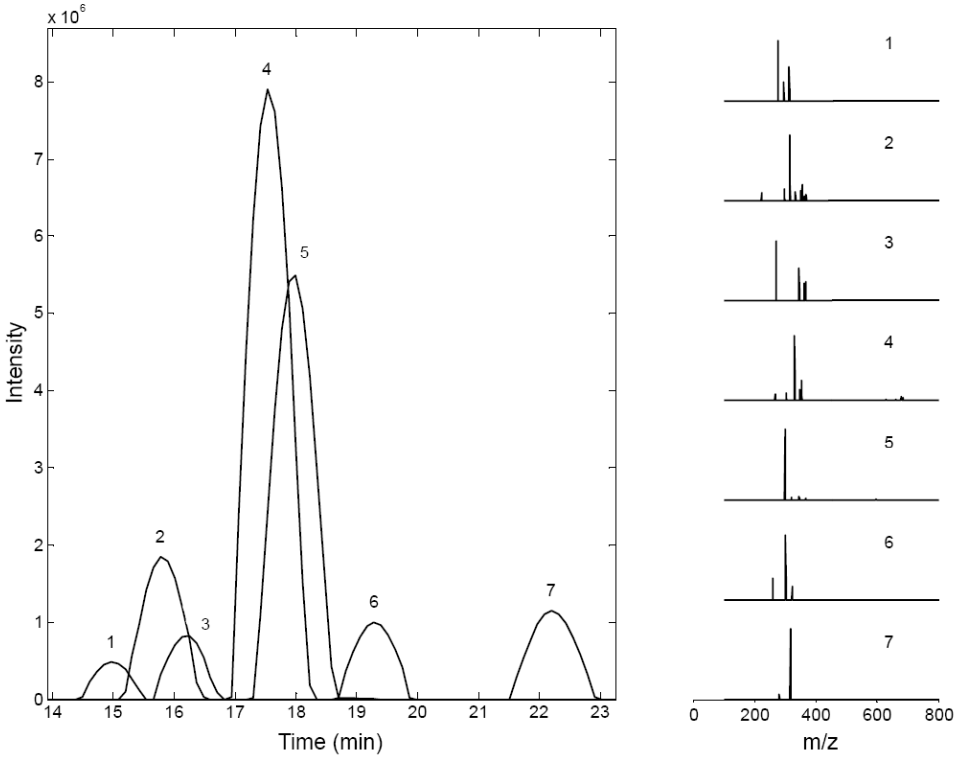

Figure 14. An example of resolved component chromatograms and respective resolved spectra of a processed experimental sample according to the proposed method in *paper III*.

If successfully executed, all components are resolved and two important estimated features are available for each component in the sample, the area of each concentration profile and the respective fingerprint in the form of a more or less unique mass spectrum. These can be utilized in the tracking of sample components between several different data sets acquired for the same sample, which is the final step towards the goal of the thesis.

7. COMPONENT TRACKING

During development of a new LC-MS method, there is a desire to achieve optimum separation of the sample constituents in minimum time. Several LC-MS experiments are thus planned to investigate how the peaks behave when the chromatographic conditions are changed. To be able to build models based on the results, each sample component needs to be identified in every experiment. This can be a difficult process, especially for samples containing a large number of components with approximately the same abundance and similar spectra. One method for component tracking is to identify all components in all samples with standards. This requires that the sample constituents are known and available as standards, however. The time needed to perform the extra analysis runs also renders this approach inappropriate.

Some of the commonly investigated chromatographic conditions include highly influential parameters such as type of column and mobile phase composition. Relatively small changes in these parameters can change the relative elution order of unidentified sample components in a more or less unpredictable manner. The identification of each component is crucial since the elution time and additional peak properties such as area, peak width and shape are important to evaluate in order to obtain the best instrumental set-up, or as input to supplemental optimisations of the method. Even with the use of a highly differentiable detector, such as the MS, a manual peak tracking approach can be extremely time inefficient.

Simple component tracking algorithms use solely the peak areas to identify components between experiments [95]. In complex samples, however, there is a greater risk that components overlap and have similar areas. Moreover, ion suppression can occur that changes the absolute response from the detector. In such cases, more sophisticated methods are required that take the spectral information into account.

7.1 Spectral similarity

Some methods were reported during the late 60s and 70s for comparing experimental spectra to previously collected spectra from databases for component identification purposes. In these cases, area information is redundant since it is not the same samples that are being compared. The construction of mass spectra databases has been in progress since then, and several large scale data bases are now available including the National Institute of Standards and Technology (NIST) [96] or Mass Bank, the official mass spectral database of The Mass Spectrometry Society of Japan [97]. The data bases are mostly based on the electron-ionisation techniques however which often are used with gas chromatography (GC). Compared to electrospray ionisation, this technique yields a different and more specific spectral fingerprint due to a more extensive breakdown of the molecules before the mass analyser.

Several of the suggested methods for comparing spectra include some sort of similarity measurement. Examples include simply counting the number of spectral entities that coincide between an arbitrary number of largest intensities [98,99], or with a theoretical template [100], by calculating the distance in space [101], by calculating the weighted ratio for each of the unknown and library spectra entities [102], by calculating the Pearson correlation coefficient [103-105], or the similar match angle [52,106-108], between the spectra together with additional match factors [109] or with weighted peak intensities [110]. Some methods function by combining several different match factors [111]. Comparisons of several of the methods showed that the match angle provided the best results when comparing spectra to a library [109] and the Pearson correlation coefficient and the match angle are closely related and often produce very similar results [112]. The use of multilayer neural networks has shown better matching abilities to library spectra when noisy and distorted experimental spectra are obtained [113]. These methodologies can also be utilized when comparing spectra from the same sample but under different chromatographic conditions. The data base then consists of the component spectra that were recently acquired, or all spectra acquired during the development of the method in progress. To fully utilize the selectiveness of LC-MS data, the combination of component area and spectra should be used when tracking components between different runs with the same sample, and the combined filtering, peak detection and component bunching step prior component tracking can be utilized to increase the matching abilities.

7.2 Other peak tracking methods

There are few complete component tracking methods described in the literature, all with more or less different processing and tracking strategies. Augmentation of several data sets prior to component tracking has been combined with PCA and a key set of pure spectra [114] and ITTFA [115]. Some strategies only involve the proton adduct in the spectra for each component for identification and tracking [116]. Another method developed for tracking components where the sample is different between each run uses the first derivative of smoothed (filtered) XICs to locate peaks and then bunches those within the same retention time window [52]. Similar components are grouped with the same match angle. The Pearson correlation coefficient has also been used to correlate the shape of chromatographic peaks between different samples in selected ion monitoring mode (i.e only one mass channel are monitored from the MS) [117]. A method referred to as alternative moving window factor analysis (AMWFA) has been proposed that is capable of extracting pure components with the information from two systems [118].

In *paper III*, a fully automatic and comprehensive component tracking method was developed, one that for each experiment performs GSD filtering, peak detection and component bunching of the data as previously explained. Relevant components are then compared between data sets for similarity regarding matching of spectral fingerprint and area of the component chromatograms. No *a priori* information about the number of components in the sample is required. It became evident from previous attempts, both by others and our own (described in the previous chapter), that perfectly resolved components would be difficult to achieve in an automatic fashion for any type of data set. Therefore, the developed tracking algorithm intentionally allowed some degree of underestimation of the actual number of components in the estimated components in the previous step. When the best match has been established between two components, only their common spectra are stored together with the respective updated component chromatograms. The residuals are then reused as potential new components during the progress of the algorithm. By using this iterative approach, smaller components embedded in larger ones can be matched and tracked. The process is exemplified in **Fig. 15** where a co-eluting peak cluster containing four components is simulated in three data sets that are shown in the first row. The components are difficult to discern, but the algorithm manages to initially estimate that there are at least two of them with the implemented PCVG based method in the two rightmost data sets, in contrast to the leftmost data set (middle row). Due to the iterative

approach when comparing the similarity of the initially guessed component spectra (not shown) together with its area, and since each component happens to be differentiated from the rest when combining the resolution from all data sets, the algorithm is capable to detect all four components in the respective data set (lowest row).

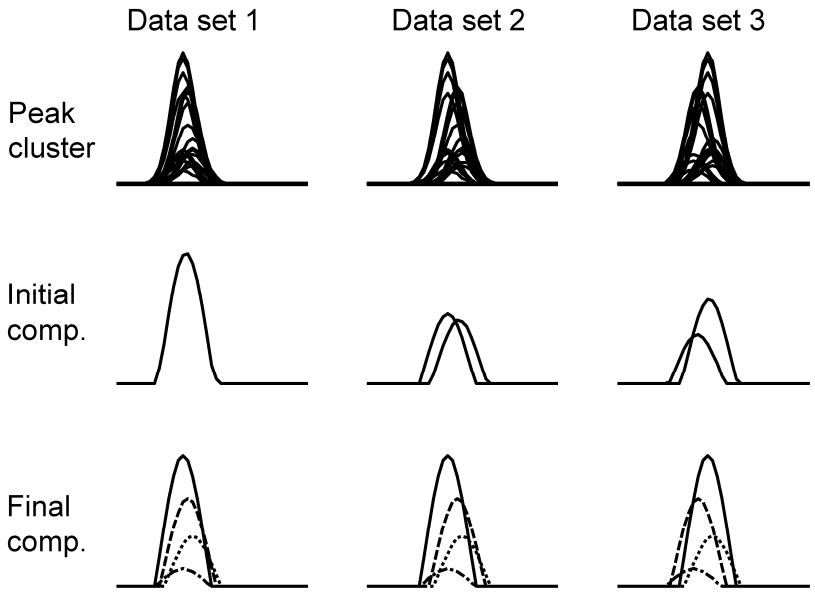

Figure 15. Simulated peak cluster of four sample components in three data sets (top row). Algorithm-generated initial component chromatograms (middle row) and algorithm-suggested final component chromatograms (bottom row).

The score of the proposed best matches, based on a weighted score from the match angle and Pearson correlation coefficient in combination with the apparent area, are listed together with the other candidates so that components with a similarity close to the top candidate can be more closely examined.

As a support, typical ions of adducts, dimers and neutral losses commonly occurring or specified by the user can be shown for the matched components, similar to the method implemented by Görlach [119]. This is performed by another matching routine where each noise free component spectra are compared with a template containing the specific spectral entities of the selected ions relative to the quasi-molecular ion ($[M+H]^+$ in positive mode). By matching with the template to each of the spectral entities of the component, the match where the most of the component spectra are explained by the different related ions becomes the

suggested protonated molecular ion and every adduct, dimer and neutral loss having a common entity to the component spectra are listed.

8 DRUG IMPURITY PROFILING

Drug impurity profiling is a generic term for the detection and identification of the impurities found in a pharmaceutical drug, originating from degradation of the drug itself during storage or from starting materials, side reactions, solvent residues during manufacturing, container materials or other possible contaminants [120]. Many pharmaceutical drug developers follow the international conference on harmonisation of technical requirements for registration of pharmaceuticals for human use (ICH) quality guidelines regarding stability testing, analysis and intake [121-126]. A review of ICH and the different guidelines can be found in [127] and a brief summary of guideline Q3A-C can be found in [128]. All impurities should be reported, but there are thresholds based on the amount of the impurity relative the amount of the active pharmaceutical ingredient whether the impurity should be identified or not (in both cases the amount needs to be established) according to **Table 1**. If unusually potent or toxic impurities are expected, identification below the threshold is advised.

Table 1. The thresholds according to ICH guidelines [124].

Maximum daily dose	Reporting threshold	Identification threshold	Qualification threshold
≤ 2g/day	0.05 %	0.10 % or 1.0 mg/day intake (whichever is lower)	0.15 % or 1.0 mg/day intake (whichever is lower)
> 2g/day	0.03 %	0.05 %	0.05 %

Prior to analysis, the pharmaceutical ingredients are dissolved and undergo accelerated stress tests in form of intense light, heat and humidity treatments, and different pH are also tested according to the ICH guidelines. Method development of the analysis of the contaminants is often performed with LC-UV-MS because of its good separation abilities and high sensitivity. During several of the steps in the process, the methods described above for detecting, bunching and tracking peaks can be utilised to aid the analyst throughout the process. In **Fig. 16**, the workflow of a typical method development is shown, identifying positions where data evaluation is required.

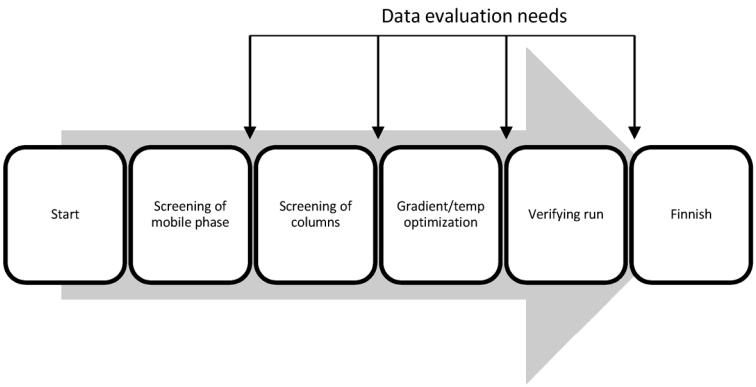

Figure 16. A typical workflow during method development of LC-UV-MS.

8.1 Choice of mobile phase and composition

As already mentioned, the aim during method development is often to find the optimal separation conditions where all sample components give rise to symmetric peaks that are baseline separated within a reasonable analysis time. Several parameters can be varied to reach this goal, each with associated practical difficulties [129]. It is common to begin the process by choosing the type of organic modifier and selecting an appropriate composition for pH, buffer and additives depending of the level of prior knowledge of the tested drug substance (e.g. pK_a). Retention generally decreases when the part of organic modifier increases in the mobile phase. If the drug and degradation products have acidic or basic properties, the pH of the mobile phase can have a large impact on the retention and selectivity. A charged organic molecule generally has a much lower affinity to the stationary phase compared to its uncharged form. The type of buffer and additive often has an impact on the shape of the peaks. If no *a priori* information is available, a screening with few experiments is performed. The purpose with this step is to find buffers that produce a good peak shape as well as sufficient retention.

8.2 Column optimisation

The type of stationary phase in the column generally has a large effect on selectivity. A tremendous selection of columns is available from different vendors, who all claim to provide the best solution for your chromatographic needs. In addition, column parameters such as length and inner diameter can influence the peak shapes considerably, but do not generally change the relative elution order (at least not for isocratic elution). To select the most appropriate column for the

current sample, some experience from similar trials can be of great assistance. Commonly, several columns are screened despite prior knowledge through a series of experiments. The selection of candidate columns can be further aided by the use of data bases where columns and their efficiency on different types of chemical groups have been analysed with PCA [130,131]. These databases have been developed so that the analyst can screen columns with stationary phases spanning a lot of the available retention and selectivity space. A column-switching valve is often utilized that is capable of automatically switching to the next column after adequate equilibrium time and after the sample has been chromatographed on the previous one (**Fig. 17**). The automatic column-switching strategy enables unsupervised over-night analysis and generates a huge amount of data that are difficult to discern and organise manually. Automated peak detection and component tracking methods can in these cases be utilized to obtain a quick and helpful guidance to the analyst.

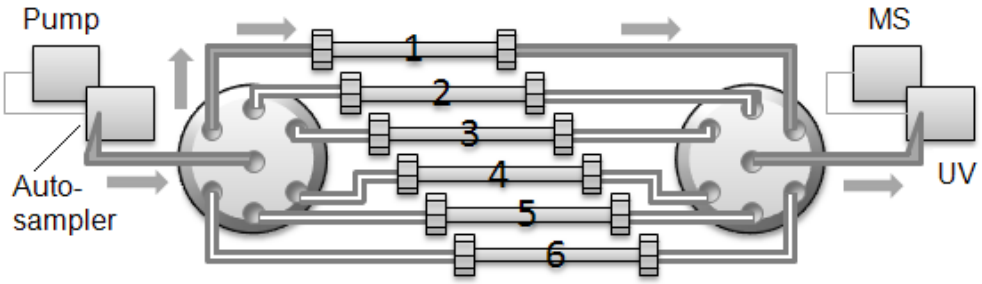

Figure 17. Schematics of a column-switching set-up capable to use six columns.

Using different mobile and stationary phase compositions can also cause different adducts to form from the same sample, thereby inducing difficulties in spectral comparison methods.

In *paper III*, the proposed method for component tracking is exemplified for experimental data acquired from a total of six different columns. The sample was an actual drug exposed to accelerated aging and was further spiked with the sample typical contaminants. The acquired data sets posed a great challenge to the algorithms since they contained a large number of false positives due to both high intensity noise spikes and missing values in low intensity chromatograms. Despite this, the method was able to accurately detect and track components in the region

of the defined thresholds in **Table 1**. The number of tracked components depends on the noise level and whether or not all components are eluting during the selected time interval. Naturally, components containing few spectral entities and low S/N ratio are more difficult to track correctly due to a lower selectiveness obtained both by a relatively larger difference in the peak area of the components and a lower number of spectral entities to compare. This is more evident when tracking components through several data sets simultaneously. With the proposed method, several components that were difficult or impossible to discern from the TIC or BPC representations could be detected in all data sets. Many of these could have remained undetected without the proposed method and their simultaneous presence in all data sets suggest that these components are actually of chemical significance. In **Fig. 18**, reconstructed ion chromatograms representing the sum of the XICs for some selected components from three of the data sets in *paper III* are shown in an interval together with an indication of the movement of the peaks. The irregular retention behaviour and a few co-eluting sites presents a great challenge to anyone attempting to perform this manually.

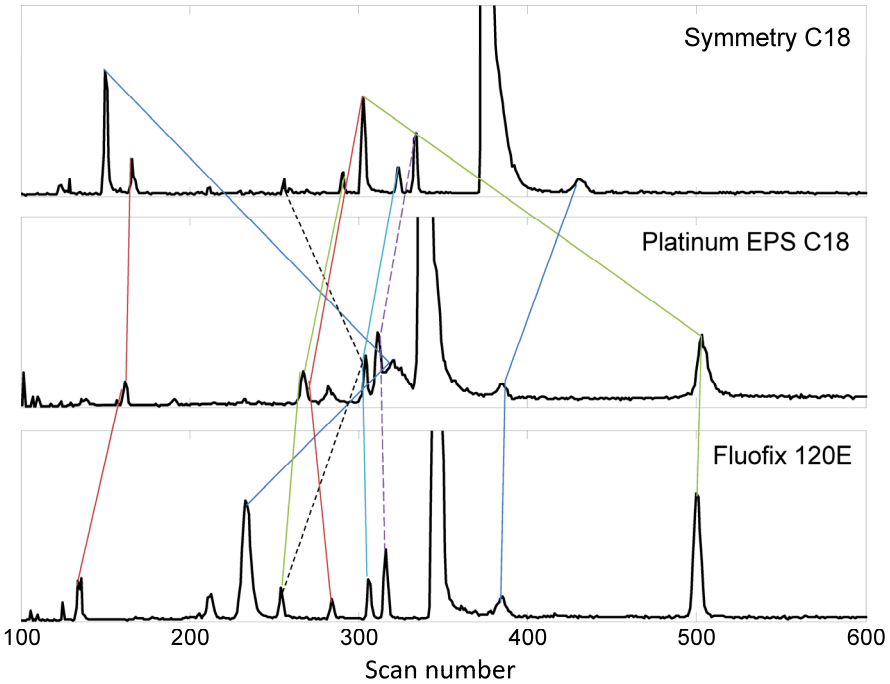

Figure 18. Truncated reconstructed ion chromatograms from the same sample acquired with three different columns. The peak movements of some selected peaks are marked.

8.3 Temperature and gradient time optimisation

When optimal column and mobile phase composition has been established, separation can be further improved by optimizing the column temperature and gradient slope. The working column temperature interval is relatively narrow in LC and is often varied to mainly control selectivity, but can also affect retention time. A higher temperature in the column fastens the equilibrium between stationary phase and analyte resulting in a more rapid elution. Different analytes are affected differently.

A linear increase of organic modifier in the mobile phase accelerates the analytes through the column. The main purpose of gradient elution is to allow the least retainable analytes to be sufficiently separated in the beginning with a low level of organic modifier, whereas the most retainable analytes will elute at an acceptable total analysis time.

The effects that the temperature and gradient have on the system are often linear and can be modelled and predicted by studying the behaviour from a set of trials on the same sample. Complex models can be used for comprehensive analysis whereas simple models are sufficient for obtaining decent retention, selectivity and peak shape predictions. With a good model, the behaviour of a system can be predicted by evaluating a set of designed experiments without the need of additional experiments.

After the designed experiments are evaluated, typically by some optimization program, the optimal conditions can be simulated and are often then verified by a new experiment. During the evaluation, properties of the component peaks are needed for optimisation which means that only components present in all experiments are applicative. During this step, a fully automatic non-user-interfered peak detection and tracking method, capable of tracking components present in all data sets with any intensity and extracting some of the required peak properties, would be of great help. The results from this step can also be used further in the tracking of the verifying experiment(s).

A strategy for the automated tracking of sample constituents during the optimisation of temperature and gradient time and verification on the simulated data is presented in *paper IV*. Here, a sample containing fluocinolone acetonide and an unknown amount of degradation products is used. The mixture was first acquired on the LC-MS instrument at three different levels of gradient steepness and two different column temperatures for building simple retention- and peak appearance-models. Relevant components could successfully be tracked in all six

data sets and their properties after data processing could be gathered in an automatic manner, minimising the time needed for manual work. In **Fig. 19**, the retention times for some automatically tracked components are shown for the six model data sets. A decreased temperature and a shallow gradient generally improved the resolution but also increased the analysis time.

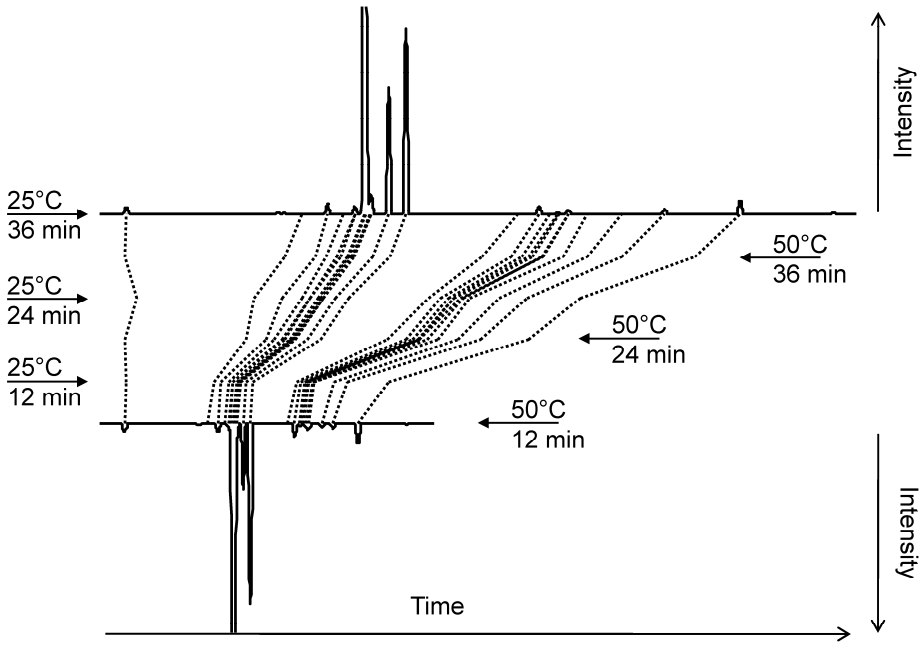

Figure 19. The change in retention times of detected and tracked components in the six model data sets acquired at different temperature (25 - 50°C) and gradient steepness (5 – 95 % ACN during 12 – 36 minutes). The two TICs (lower is upside down) are truncated at 50 % of the main signal.

Excellent retention models and decent peak appearance models could be achieved when compared to the experiment made for model verification. The obtained minimum resolution between the selected and optimized component peaks differed somewhat from the calculated value but the raw and processed data showed similar resolution values. Many of the tracked components contained m/z ratios in the same region of the expected ratio of the original substance, and could be found in both the model and the verifying data sets at levels of interest for the current application.

9 CONCLUDING REMARKS AND FURTHER PERSPECTIVES

Separation of unknown mixtures during method development and routine analytical work requires functional instruments, knowledge, time and effort as we want to learn more about the composition in our samples. Highly automated LC-MS systems are routinely loaded with more and more complex samples where an incredibly large amount of data can be acquired more or less unsupervised. Interpretation and implementation of the resulting data are, however, seldom fully automated and therefore require manual or semi-manual strategies. During this process, finding and structuring data often becomes inefficient and subjective. As a consequence, in a wider perspective, the progress in science becomes delayed.

Development of newer chromatography systems such as the fast, sub 2μm particle with high back pressure, ultra-HPLC systems have further decreased the time needed for analysis and acquiring the data sets. Focus should perhaps now be shifted more to the data evaluation techniques as these begin to be just as important for the total analysis throughput.

The proposed strategies were developed for a typical sample consisting of a pharmaceutical drug and its degradation products, even though other types of data sets were sometimes used. It is believed that a similar strategy can be used for data acquired from different separation methods such as gas chromatography (GC) or capillary electrophoresis (CE) as well as with widely different samples. The methods could for example be used to aid the field of proteomics/metabolomics/lipidomics, where determination of the existence and levels of proteins or metabolites in complex matrices is a common task [132]. A few tweaks of the involved algorithms can be necessary however, such as increasing the resolution (with the cost of lower S/N) during peak detection since these samples often have a large number of constituents. Moreover, samples from different origins give rise to different component peak areas which renders the contribution of areas to the similarity measurement unfeasible when tracking. Spectral similarity alone could be sufficient or an extra weight could be introduced to the area contribution.

Moreover, data from the UV detector, which is often used as a complement to the MS could be incorporated, both to include non-ionisable components that cannot be detected by the MS and extract complementary information that can be incorporated to increase the differentiable power for the sample components. This has earlier shown to improve curve resolution [80,81]. Alignment of the data from the different detectors would be required although this should not present a

problem, but similar peak shape seems also to be of importance for the same component, which can be more troublesome.

There is a great number and wide variety of strategies explained in the literature to automatically or semi-automatically aid the analyst during data evaluation. Naturally, these are often tailored for the developer's specific problem formulation and sometimes readily available for anyone with Internet access. This is good for experts that want to use and tweak the underlying algorithms for their own purposes in contrast to what often is possible with the black-box vendor programs available. For a less experienced user that just wants to try the ideas on their own data, however, there is often a myriad of hurdles to overcome in the form of specific knowledge of different required operating systems, coding languages and parameter settings. Moreover, one's data should often be stored in one of many different ways in order to be compatible. To become more publicly accessible, all published algorithms in the field should ultimately be collected and converted to a single program with an open source code. This way, each scientist can contribute with his or her own algorithms that anyone can easily access and test. The different algorithms can function as seeds for new ideas as further development in this field are essential. Successful applications and shortcomings could be reported by any user.

Some particular problems are not easily solved with an algorithm. Methods such as artificial neural networks (ANNs) exist that attempt to mimic the human brain on a smaller scale, as the brain has extraordinary advantages in pattern recognition and logical processing which are some of the useful properties essential for all stages in LC-MS data interpretation. These attributes cannot yet be fully mimicked with a computer, however. The development of visualization and refinement strategies would thus be a natural next step for the proposed methods, where ambiguous results could rapidly be made clear by an everyday experienced analytical chemist. The computer program should then in a few steps be able to produce sophisticated suggestions by which the user can make decisions. Even though the true solution for several ambiguous results can be obvious for an experienced user and not for an algorithm making decisions on mathematical models, incidents do occur, however, where the true solution cannot be decided by either man or machine.

ACKNOWLEDGEMENTS

My sincere gratitude to my head supervisor Dan Bylund for believing in me, for his guidance and for being calm in hectic situations, my associate supervisors Patrik Petersson and Bengt-Olof Axelsson for their enthusiasm and invaluable contributions. I also want to acknowledge my former supervisors Ulla Lundström and Magnus Jörntén-Karlsson.

The rest of the group of analytical and soil chemistry, Tara, Madde and Sara N, the optimal office roommate, and the former members Sara H, Sofia E, Per-Erik M and Jenny V are acknowledged for scientific discussions and for being nice persons.

The colleagues at the Department of Natural Sciences, Engineering and Mathematics (NAT) gets huge kudos for making this place a nice place to work at. Extra kudos goes to Anna H, Håkan N and Torborg J for all practical help during the years.

Thanks to all former and present fika buddies and friends for enjoyable discussions and that you never believe in my stories.

I also wish to give gratitude to my old study pals during my graduate studies and my training buddy's at Helex which literally keep me on my toes.

Thanks to Per Edström for a critical inspection of the thesis and to Matt Richardson for suggesting linguistic improvements.

My gratitude also goes to my parents, for guidance and support during the early years of my life and to my brother and sister with families and other relatives and friends which have been neglected during the most intense parts of this work. I hope to be able to spend some more time with all of you soon.

Susanne, thanks for being my buffer, catalyst and very best friend. I love you.

I'd also like to thank the word "component" (602) and "peak" (1080 instances!) for their existence, because without their presence the number of blank pages in this thesis would be considerable larger. Moreover, this corresponds somewhat to what is typically present in the LC-MS data sets investigated herein. One may then wonder, however, if the rest of the text is to be considered as noise…

REFERENCES

[1] N. Lundell, K. Markides, Journal of Chromatography 639 (1993) 117.
[2] D. Guillarme, J. Ruta, S. Rudaz, J.L. Veuthey, Analytical and Bioanalytical Chemistry 397 1069.
[3] D.T.T. Nguyen, D. Guillarme, S. Rudaz, J.L. Veuthey, Journal of Chromatography A 1128 (2006) 105.
[4] D.A. Skoog, F.J. Holler, T.A. Nieman, Principles of Instrumental Analysis, Brooks/Cole, 1997.
[5] P.D. Wentzell, C.D. Brown, Signal Processing in Analytical Chemistry, John Wiley & Sons Ltd Cichester, 2000.
[6] C.D. Brown, P.D. Wentzell, Journal of Chemometrics 13 (1999) 133.
[7] W.E. Wallace, A.J. Kearsley, C.M. Guttman, Analytical Chemistry 76 (2004) 2446.
[8] M.H.J. van Rijswick, Chromatographia 7 (1974) 491.
[9] B. Vandenbogaert, H.F.M. Boelens, H.C. Smit, Analytica Chimica Acta 274 (1993) 71.
[10] B. Vandenbogaert, H.F.M. Boelens, H.C. Smit, Analytica Chimica Acta 274 (1993) 87.
[11] A.V. Oppenheim, R.W. Schafer, J.B. Buck, Discrete-Time Signal Processing, Prentice Hall, 1999.
[12] S.W. Smith, The Scientist and Engineer's Guide to Digital Signal Processing, California Technical Pub.; 1st edition (1997), 1997.
[13] A. Solis, M. Rex, A.D. Campiglia, P. Sojo, Electrophoresis 28 (2007) 1181.
[14] R. Danielsson, D. Bylund, K.E. Markides, Analytica Chimica Acta 454 (2002) 167.
[15] H.H. Madden, Analytical Chemistry 50 (1978) 1383.
[16] A. Savitzky, M.J.E. Golay, Analytical Chemistry 36 (1964) 1627.
[17] P. Barak, Analytical Chemistry 67 (1995) 2758.
[18] M. Jakubowska, W.W. Kubiak, Analytica Chimica Acta 512 (2004) 241.
[19] G. Vivo-Truyols, P.J. Schoenmakers, Analytical Chemistry 78 (2006) 4598.
[20] J.W. Luo, K. Ying, J. Bai, Signal Processing 85 (2005) 1429.
[21] C. Chinrungrueng, in 2003 Ieee International Conference on Systems, Man and Cybernetics, Vols 1-5, Conference Proceedings, 2003, p. 690.
[22] Y.F. Shen, M.L. Lee, Analytical Chemistry 70 (1998) 3853.
[23] W. Windig, J.M. Phalp, A.W. Payne, Analytical Chemistry 68 (1996) 3602.
[24] W. Windig, W.F. Smith, Journal of Chromatography A 1158 (2007) 251.
[25] W. Windig, W.F. Smith, W.F. Nichols, Analytica Chimica Acta 446 (2001) 467.

[26] X.G. Shao, A.K.M. Leung, F.T. Chau, Accounts of Chemical Research 36 (2003) 276.
[27] V.J. Barclay, R.F. Bonner, I.P. Hamilton, Analytical Chemistry 69 (1997) 78.
[28] S. Cappadona, F. Levander, M. Jansson, P. James, S. Cerutti, L. Pattini, Analytical Chemistry 80 (2008) 4960.
[29] E. Lange, C. Gropl, K. Reinert, O. Kolbacher, A. Hildebrandt, in Pacific Symposium on Biocomputing, 2006, p. 243.
[30] A.K.M. Leung, F.T. Chau, J.B. Gao, Chemometrics and Intelligent Laboratory Systems 43 (1998) 165.
[31] E. Mostacci, C. Truntzer, H. Cardot, P. Ducoroy, Proteomics 10 2564.
[32] X.G. Shao, W.S. Cai, Z.X. Pan, Chemometrics and Intelligent Laboratory Systems 45 (1999) 249.
[33] X.G. Shao, C.X. Ma, Chemometrics and Intelligent Laboratory Systems 69 (2003) 157.
[34] B. Walczak, D.L. Massart, Chemometrics and Intelligent Laboratory Systems 36 (1997) 81.
[35] A. Ghosh, R.J. Anderegg, Analytical Chemistry 61 (1989) 73.
[36] W.G. Pool, J.W. deLeeuw, B. vandeGraaf, Journal of Mass Spectrometry 31 (1996) 509.
[37] C.M. Fleming, B.R. Kowalski, A. Apffel, W.S. Hancock, Journal of Chromatography A 849 (1999) 71.
[38] D.C. Muddiman, B.M. Huang, G.A. Anderson, A. Rockwood, S.A. Hofstadler, M.S. WeirLipton, A. Proctor, Q.Y. Wu, R.D. Smith, Journal of Chromatography A 771 (1997) 1.
[39] M.A. Elliott, G.A. Walter, A. Swift, K. Vandenborne, J.C. Schotland, J.S. Leigh, Magnetic Resonance in Medicine 41 (1999) 450.
[40] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Analytical Chemistry 78 (2006) 779.
[41] V.P. Andreev, T. Rejtar, H.S. Chen, E.V. Moskovets, A.R. Ivanov, B.L. Karger, Analytical Chemistry 75 (2003) 6314.
[42] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M.C. Hung, H.M. Kuerer, Proteomics 5 (2005) 4107.
[43] C. Yang, Z.Y. He, W.C. Yu, Bmc Bioinformatics 10 (2009).
[44] F. Gan, G.H. Ruan, J.Y. Mo, Chemometrics and Intelligent Laboratory Systems 82 (2006) 59.
[45] P.H.C. Eilers, Analytical Chemistry 75 (2003) 3631.
[46] R. Sanchez-Ponce, F.P. Guengerich, Analytical Chemistry 79 (2007) 3355.
[47] S. Ullsten, R. Danielsson, D. Backstrom, P. Sjoberg, J. Bergquist, Journal of Chromatography A 1117 (2006) 87.

[48] P. Jonsson, S.J. Bruce, T. Moritz, J. Trygg, M. Sjostrom, R. Plumb, J. Granger, E. Maibaum, J.K. Nicholson, E. Holmes, H. Antti, Analyst 130 (2005) 701.
[49] H. Rosing, W.Y. Man, E. Doyle, A. Bult, J.H. Beijnen, Journal of Liquid Chromatography & Related Technologies 23 (2000) 329.
[50] S.E. Stein, Journal of the American Society for Mass Spectrometry 10 (1999) 770.
[51] Analytical Chemistry 52 (1980) 2242.
[52] S.J. Dixon, R.G. Brereton, H.A. Soini, M.V. Novotny, D.J. Penn, Journal of Chemometrics 20 (2006) 325.
[53] A. Antoniadis, J. Bigot, S. Lambert-Lacroix, F. Letue, Current Analytical Chemistry 3 (2007) 127.
[54] E.J. Breen, F.G. Hopwood, K.L. Williams, M.R. Wilkins, Electrophoresis 21 (2000) 2243.
[55] P. Du, W.A. Kibbe, S.M. Lin, Bioinformatics 22 (2006) 2059.
[56] K.H. Jarman, D.S. Daly, K.K. Anderson, K.L. Wahl, Chemometrics and Intelligent Laboratory Systems 69 (2003) 61.
[57] M. Katajamaa, M. Oresic, Journal of Chromatography A 1158 (2007) 318.
[58] A.J. Kearsley, W.E. Wallace, J. Bernal, C.M. Guttman, Applied Mathematics Letters 18 (2005) 1412.
[59] T.W. Randolph, Y. Yasui, Biometrics 62 (2006) 589.
[60] C.S. Tan, A. Ploner, A. Quandt, J. Lehtio, Y. Pawitan, Bioinformatics 22 (2006) 1515.
[61] O. Schulz-Trieglaff, R. Hussong, C. Gropl, A. Hildebrandt, K. Reinert, in T. Speed, H. Huang (Editors), Research in Computational Molecular Biology, Proceedings, 2007, p. 473.
[62] C.A. Hastings, S.M. Norton, S. Roy, Rapid Communications in Mass Spectrometry 16 (2002) 462.
[63] R. Stolt, R.J.O. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, S.P. Jacobsson, Analytical Chemistry 78 (2006) 975.
[64] K. Lan, J.W. Jorgenson, Analytical Chemistry 71 (1999) 709.
[65] D. Bylund, R. Danielsson, K.E. Markides, Journal of Chromatography A 915 (2001) 43.
[66] R. Tauler, D. Barcelo, Trac-Trends in Analytical Chemistry 12 (1993) 319.
[67] R. Tauler, B. Kowalski, S. Fleming, Analytical Chemistry 65 (1993) 2040.
[68] P.J. Gemperline, Journal of Chemical Information and Computer Sciences 24 (1984) 206.
[69] B.G.M. Vandeginste, W. Derks, G. Kateman, Analytica Chimica Acta 173 (1985) 253.

[70] A. de Juan, R. Tauler, Critical Reviews in Analytical Chemistry 36 (2006) 163.
[71] A.K. Smilde, H.C.J. Hoefsloot, H.A.L. Kiers, S. Bijlsma, H.F.M. Boelens, Journal of Chemometrics 15 (2001) 405.
[72] R. Tauler, Chemometrics and Intelligent Laboratory Systems 30 (1995) 133.
[73] E.R. Malinowski, Journal of Chemometrics 6 (1992) 29.
[74] R. Manne, H.L. Shen, Y.Z. Liang, Chemometrics and Intelligent Laboratory Systems 45 (1999) 171.
[75] I.H.M. van Stokkum, K.M. Mullen, V.V. Mihaleva, Chemometrics and Intelligent Laboratory Systems 95 (2009) 150.
[76] J.H. Jiang, Y.Z. Liang, Y. Ozaki, Chemometrics and Intelligent Laboratory Systems 71 (2004) 1.
[77] P.V. van Zomeren, H. Darwinkel, P.M.J. Coenegracht, G.J. de Jong, Analytica Chimica Acta 487 (2003) 155.
[78] E. Pere-Trepat, S. Lacorte, R. Tauler, Journal of Chromatography A 1096 (2005) 111.
[79] R. Tauler, A. Izquierdoridorsa, R. Gargallo, E. Casassas, Chemometrics and Intelligent Laboratory Systems 27 (1995) 163.
[80] C. Bessant, R.G. Brereton, S. Dunkerley, Analyst 124 (1999) 1733.
[81] E. Pere-Trepat, R. Tauler, Journal of Chromatography A 1131 (2006) 85.
[82] P.V. van Zomeren, H.J. Metting, P.M.J. Coenegracht, G.J. de Jong, Journal of Chromatography A 1096 (2005) 165.
[83] J. Forshed, H. Idborg, S.P. Jacobsson, Chemometrics and Intelligent Laboratory Systems 85 (2007) 102.
[84] H.T. Gao, T.H. Li, K. Chen, S.F. Lin, Talanta 68 (2006) 542.
[85] R. Gargallo, R. Tauler, F. CuestaSanchez, D.L. Massart, Trac-Trends in Analytical Chemistry 15 (1996) 279.
[86] M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi, M. Maeder, Journal of Chemometrics 20 (2006) 302.
[87] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbuhler, Talanta 33 (1986) 943.
[88] M. Maeder, Analytical Chemistry 59 (1987) 527.
[89] H.R. Keller, D.L. Massart, Analytica Chimica Acta 246 (1991) 379.
[90] H.R. Keller, D.L. Massart, J.O. Debeer, Analytical Chemistry 65 (1993) 471.
[91] W. Windig, J. Guilment, Analytical Chemistry 63 (1991) 1425.
[92] F.C. Sanchez, V. vandenBogaert, S.C. Rutan, D.L. Massart, Chemometrics and Intelligent Laboratory Systems 34 (1996) 139.
[93] L. Xu, L.J. Tang, C.B. Cai, H.L. Wu, G.L. Shen, R.Q. Yu, J.H. Jiang, Analytica Chimica Acta 613 (2008) 121.
[94] G. Ivosev, L. Burton, R. Bonner, Analytical Chemistry 80 (2008) 4933.

[95] I. Molnar, Journal of Chromatography A 965 (2002) 175.
[96] S.E. Stein, NIST Chemistry WebBook, NIST Standard Reference Database Number 69, 2010.
[97] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, Journal of Mass Spectrometry 45 703.
[98] S.L. Grotch, Analytical Chemistry 42 (1970) 1214.
[99] B.A. Knock, I.C. Smith, D.E. Wright, R.G. Ridley, W. Kelly, Analytical Chemistry 42 (1970) 1516.
[100] D.W. Hill, T.M. Kertesz, D. Fontaine, R. Friedman, D.F. Grant, Analytical Chemistry 80 (2008) 5574.
[101] L.R. Crawford, J.D. Morrison, Analytical Chemistry 40 (1968) 1464.
[102] H.S. Hertz, R.A. Hites, K. Biemann, Analytical Chemistry 43 (1971) 681.
[103] B.Y. Li, Y. Hu, Y.Z. Liang, L.F. Huang, C.J. Xu, P.S. Xie, Journal of Separation Science 27 (2004) 581.
[104] W. Li, C.Q. Hu, Journal of Chromatography A 1190 (2008) 141.
[105] M.J. Sniatynski, J.C. Rogalski, M.D. Hoffman, J. Kast, Analytical Chemistry 78 (2006) 2600.
[106] S.E. Stein, Journal of the American Society for Mass Spectrometry 6 (1995) 644.
[107] M.E. Swartz, P.R. Brown, Chirality 8 (1996) 67.
[108] K.X. Wan, I. Vidavsky, M.L. Gross, Journal of the American Society for Mass Spectrometry 13 (2002) 85.
[109] S.E. Stein, D.R. Scott, Journal of the American Society for Mass Spectrometry 5 (1994) 859.
[110] W. Demuth, M. Karlovits, K. Varmuza, Analytica Chimica Acta 516 (2004) 75.
[111] K.S. Kwok, Venkatar.R, McLaffer.Fw, Journal of the American Chemical Society 95 (1973) 4185.
[112] F. Gong, B.T. Wang, F.T. Chau, Y.Z. Liang, Analytical Letters 38 (2005) 2475.
[113] C.S. Tong, K.C. Cheng, Chemometrics and Intelligent Laboratory Systems 49 (1999) 135.
[114] A. Bogomolov, M. McBrien, Analytica Chimica Acta 490 (2003) 41.
[115] P.V. van Zomeren, A. Hoogvorst, P.M.J. Coenegracht, G.J. de Jong, Analyst 129 (2004) 241.

[116] G. Xue, A.D. Bendick, R. Chen, S.S. Sekulic, Journal of Chromatography A 1050 (2004) 159.
[117] J.H. Christensen, J. Mortensen, A.B. Hansen, O. Andersen, Journal of Chromatography A 1062 (2005) 113.
[118] Z.D. Zeng, Y.Z. Liang, Y.L. Wang, X.R. Li, L.M. Liang, Q.S. Xu, C.X. Zhao, B.Y. Li, F.T. Chau, Journal of Chromatography A 1107 (2006) 273.
[119] E. Görlach, R. Richmond, Analytical Chemistry 71 (1999) 5557.
[120] N. Rahman, S.N.H. Azmi, H.F. Wu, Accreditation and Quality Assurance 11 (2006) 69.
[121] ICH Harmonised Tripartite Guideline (2003) Stability testing of new drug substances and products Q1A(R2).
[122] ICH Harmonised Tripartite Guideline (1996) Stability Testing: Photostability testing of new drug substances and products Q1B.
[123] ICH Harmonised Tripartite Guideline (2005) Validation of Analytical Procedures: Text and Methodology Q2(R1).
[124] ICH Harmonised Tripartite Guideline (2006) Impurities in New Drug Substances Q3A(R2).
[125] ICH Harmonised Tripartite Guideline (2006) Impurities in New Drug Products Q3B(R2).
[126] ICH Harmonised Tripartite Guideline (2009) Impurities: Guideline for Residual Solvents Q3C(R4).
[127] S.K. Branch, Journal of Pharmaceutical and Biomedical Analysis 38 (2005) 798.
[128] T. McGovern, D. Jacobson-Kram, Trac-Trends in Analytical Chemistry 25 (2006) 790.
[129] L.R. Snyder, Journal of Chromatography B 689 (1997) 105.
[130] M. Euerby, P. Petersson, Journal of Chromatography A 994 (2003) 13.
[131] M.R. Euerby, P. Petersson, Journal of Chromatography A 1088 (2005) 1.
[132] J. Trygg, E. Holmes, T. Lundstedt, Journal of Proteome Research 6 (2007) 469.