Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

A component tracking algorithm for accelerated and improved liquid chromatography–mass spectrometry method development

Mattias J. Fredriksson[a], Patrik Petersson[b], Bengt-Olof Axelsson[b], Dan Bylund[a,*]

[a] Mid Sweden University, Department of Natural Sciences, Engineering and Mathematics, SE-851 70 Sundsvall, Sweden
[b] AstraZeneca, Analytical Development, R&D Lund, SE-221 87 Lund, Sweden

ARTICLE INFO

Article history:
Received 14 June 2010
Received in revised form 12 October 2010
Accepted 25 October 2010
Available online 30 October 2010

Keywords:
Liquid chromatography
Mass spectrometry
Chemometrics
Peak detection
Peak tracking

ABSTRACT

A method for tracking of sample components during liquid chromatography–mass spectrometry (LC–MS) method development has been proposed. The method manages to, fully automatically and without user intervention, find the chromatographic peaks in the data sets, discriminate them to sample components and track them when the separation conditions have been changed. The algorithm utilises the resolution obtained from all considered data sets and has the ability to discriminate the non informative parts. The technique has a great sensitivity even in cases where a majority of the tracked components cannot easily be spotted by means of traditional total ion chromatogram (TIC) or base peak chromatogram (BPC) representations. The method was tested on an experimental sample using six different columns and an average of 79% of the suggested sample components could be successfully tracked at a minimum area of 0.05% of the main component in the sample. 66 components with 79–92% of the total suggested component area were able to be tracked between all data sets. The method could be used to rapidly investigate selectivity during different types of separation conditions.

1. Introduction

Keeping track of the sample components eluting from a chromatographic system under various conditions is a critical task during liquid chromatography (LC) method development. Unless the peaks are correctly assigned, erroneous retention models or conclusions will be obtained resulting in sub-optimal results and/or prolonged time and efforts required for the development process.

In the pharmaceutical industry, drug impurity profiling is an important area for which highly efficient LC methods are called for. Drug impurity profiling is a generic term for the determination of chemical entities not defined as the drug substance [1]. There are regulations regarding how these impurities must be scrutinised and controlled by the pharmaceutical industry. The statute can be summarised as being that the drug developers must be able to report the presence of and quantify all degradation products down to 0.05% of the active drug substance, whereas degradation products down to 0.1% also need to be identified structurally. The stability of the drug substance is commonly investigated by exposing the substance of interest to an accelerated ageing process generated by light, low/high pH, humidity and elevated temperature. This treatment often results in an unknown mixture of the drug substance and the degradation impurities. In addition, contaminants originating from preparation such as solvent residues, unwanted side-reactions and packaging may also be present.

To address this complicated analytical task, the LC system is often coupled on-line to both a diode array detector (DAD) and a mass spectrometer (MS) operated in full scan mode, thereby enabling the simultaneous detection of both light absorbing and ionisable compounds. After selection of a suitable pH and buffer based on the structure of the drug and/or scouting experiments, the LC conditions are often optimized in a two-step process. Firstly, columns with different selectivity are screened in order to identify a column which provides a good peak shape, gives sufficient retention and which separates the largest number of components. Another reason for this screening is to determine which components are present in the sample(s). Following this, the operating conditions for the selected column are identified, often with the focus on the gradient profile and temperature. In both of these optimization steps it is important to track the components in the generated LC-DAD–MS chromatograms. The first step is the most complicated since the peaks can move around more or less randomly due to the selectivity differences of the columns tested. Thus, the component tracking method described in this work was tested on such a data set.

The instrument set-up for the screening of several columns usually contains an automatic column switching valve, where the columns to be tested are selected to be as orthogonal in selectivity as possible. This column selection is preferably aided by a column

* Corresponding author. Tel.: +46 60 148909; fax: +46 60 148820.
E-mail address: dan.bylund@miun.se (D. Bylund).

0021-9673/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2010.10.083

characterisation database [2]. This enables a relatively straightforward, time efficient and comprehensive analysis at different separation conditions. The subsequent data analysis, however, is a major bottleneck during the method development. The peak tracking step is often performed manually, by comparing the corresponding UV and MS spectra, and since little or no information is available for the sample mixture, the task is tedious and error prone.

The most intuitive manual method for the matching of LC-DAD–MS peaks is by inspecting and comparing the peak intensities or areas beneath the graph in the UV chromatograms. Several peaks may, however, have similar sizes, which thus makes it impossible to perform the identification by mere intensity. The next step would then be to utilise the spectral information from each component in the sample mixture. Unfortunately, the signal to noise ratio of UV chromatograms and UV spectra is usually relatively poor at the 0.05% level. As a consequence, peak tracking based on these parameters becomes uncertain. An advantage associated with MS spectra is that they are often more component specific than the UV spectra, and thus easier to use in order to distinguish between different sample components. Some components in a sample mixture can, however, have very similar mass spectra (especially when soft ionization methods, such as electrospray, are applied), and in these cases the use of spectral correlation alone may provide multiple candidates for matching. A combination of intensity/area information and spectral information about the components in a sample would thus be advantageous for the accurate matching of chromatographic peaks between data sets.

Comparing MS spectra by an automatic approach is also subject to difficulties. Interfering high intensity background signals and noise originating from uncontrollable experimental variances can blur the data and suppress the similarities between matching sample components [3]. It is thus important that the spectral information for each component in the sample is preserved regardless of the separation conditions. Present data sets commonly contain thousands of mass channels, so some processing is required to reduce the otherwise significant number of possible matches that require to be tested. In addition, several decently separated components in the chromatogram generated from one set of separation conditions can partially, or totally, co-elute (or simply not elute at all) during the next selected conditions. To be able to efficiently scrutinise the automatic results, it is thus important that co-eluted components can be treated individually and that no match is registered if a component does not have a corresponding component in the other data set.

Automated spectral correlation techniques were developed extensively during the late 1960s and 1970s for different analytical techniques [4–12], but have also received more attention recently [13–33]. Rasmussen and Isenhour reported in 1979 that there are basically two main elements in action when spectra are to be matched, namely the data encoding method, where the parts of the original spectra to be used in the comparison are determined, and the comparison algorithm, which describes how spectra are to be compared [4]. They also mention the use of prefilters, which are used to rule out improbable candidates at an early stage in order to increase performance. Albeit the instrumental and computational powers have improved since then, the current approach is fundamentally based on these elements. Few investigations consisting of a fully automatic approach for tracking components from the same sample but during different separation conditions have been reported in the literature. Swartz and Brown describe a method where one of their experiments is used as a reference library in order to compare, track and check the purity of the sample components when some chromatographic conditions are changed [15]. Mutual peak matching (MAP) is a method for LC-DAD proposed by Bogomolov and McBrien which adjoins data from the same mixture sampled at various conditions and a key set of spectra are used for matching [29]. van Zomeren and co-workers used the method of augmented iterative target transformation factor analysis to both resolve overlapping peaks and to track them between the different data sets simultaneously [30]. COMET, which is an abbreviation for comprehensive orthogonal method evaluation, is reported as being a fully automated method for tracking LC–MS components during impurity profiling in a paper by Xue et al. [31]. Dixon et al. proposed a method for peak detection and matching of pre-processed GC–MS data. Mass channels are examined individually for peaks which are then grouped into components and compared by a similarity measurement [32]. Zeng et al. proposed a similar method to the current one in which the information from two systems is utilised in order to determine the number of common components between them and to extract pure spectra by using an alternative moving window factor analysis (AMWFA) [33].

In this work, we present a peak tracking method which combines strategies and ideas from previously developed methods for various tasks. These were refined to various degrees to better harmonize to our requirements. The aim was to fully automatically track and highlight even the smallest components in a sample while simultaneously disregard irrelevant information. The method starts from individual unprocessed raw data and no specific knowledge about its structure is required. It is capable to enhance and extract the informative peaks by automatically determine the essential algorithm parameters. The outcome from this peak detection is a noise- and baseline-free reconstruction in which peaks belonging to the same sample component are further evaluated by the algorithm. These components are then compared for similarity in terms of relative spectral and total intensity between the investigated data sets. The algorithm provides new data for each included data set containing the component chromatogram and corresponding spectrum together with a list of selected and rejected matches and is capable to automatically discriminate false positives and other artefacts.

2. Theory

A flowchart of the new component tracking algorithm can be viewed in Fig. 1. The most important steps will be further explained in the sections below. For a more complete understanding, we also refer to supplementary material and to the references given.

2.1. LC–MS data

A single data set obtained from an LC–MS instrument which scans several pre defined mass to charge (m/z) ratios discretely over time can be regarded as an $m \times n$ matrix, $\mathbf{D}$, with m rows (m/z) and n columns (time). $\mathbf{D}$ can be separated into a matrix $\mathbf{A}$ which contains the useful analytical signals (the peaks), $\mathbf{B}$, which contains the low frequency signals (the background), and $\mathbf{E}$ which contains the high frequency signals (the noise) according to Eq. (1).

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

If we reduce $\mathbf{B}$ and $\mathbf{E}$, by applying a peak detection algorithm in the time domain and only store the detected peaks, each column, n, in the resulting data matrix, $\mathbf{D}_{\mathrm{pd}}$, will obtain a mass spectrum with fewer but more characteristic mass ions corresponding to the informative parts of the original data set. $\mathbf{D}_{\mathrm{pd}}$ can in turn be deconvoluted by Eq. (2) into $j \times n$ concentration profiles $\mathbf{C}$ and $j \times m$ mass spectra $\mathbf{S}$, where j is the number of components in the sample mixture.

$$\mathbf{D}_{\mathrm{pd}} = \mathbf{S}^{\mathrm{T}}\mathbf{C} + \mathbf{R} \quad (2)$$

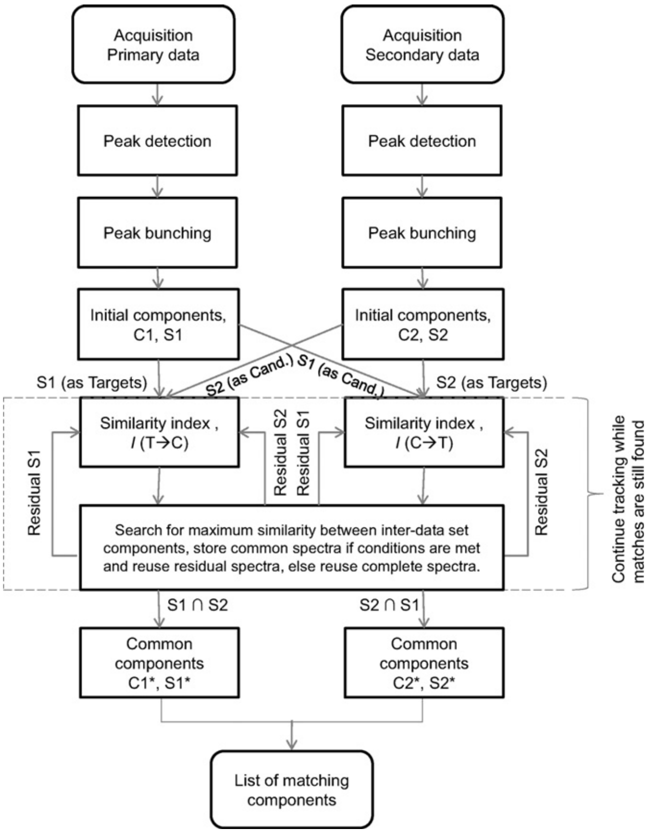
Acquisition Primary data
Acquisition Secondary data
Peak detection
Peak detection
Peak bunching
Peak bunching
Initial components, C1, S1
Initial components, C2, S2
S1 (as Targets)
S2 (as Cand.)
S1 (as Cand.)
S2 (as Targets)
Similarity index , I (T→C)
Similarity index , I (C→T)
Residual S1
Residual S2
Residual S1
Residual S2
Search for maximum similarity between inter-data set components, store common spectra if conditions are met and reuse residual spectra, else reuse complete spectra.
Continue tracking while matches are still found
S1 ∩ S2
S2 ∩ S1
Common components C1*, S1*
Common components C2*, S2*
List of matching components

Fig. 1. Flowchart of the component matching algorithm.

In this case, **R** is the residual error and T denotes the transpose of the matrix **S**. The deconvolution can be seen as the bunching of all the chromatographic peaks in $\mathbf{D}_{\mathrm{pd}}$ which have approximately the same retention time (t_{R}) into one component specific elution profile in **C** with a corresponding spectrum in **S**. If the deconvolution is successful, **R** is minimised and information about the retention times, spectra and the number of components in the sample mixture is readily available for further processing. The bunching procedure is often, however, problematic since components may co-elute. Many methods for resolving overlapped peaks have been developed throughout previous years. They are referred to as multivariate curve resolution (MCR) techniques, which almost exclusively solve the problem by searching for linear combinations of the data.

If genuine components are obtained consisting of single, Gaussian shaped concentration profiles with the correct intensities and with the corresponding pure spectra, the components can act as a reference library for the data sets from the same sample but which have been acquired during different separation conditions. The library spectra could be used to find positions in time in the new data set where a similar spectrum is obtained by calculating some sort of similarity index. In addition, the second data set should be deconvoluted into concentration profiles and spectra in the same manner in order to reduce the interference from the background and noise and to decrease the calculation time during the matching procedure by reducing the number of possible matching candidates. Pure components are, however, difficult to obtain and to utilise the resolution power from two separation conditions, the components should be compared both forward and backward so that components that are unable to be resolved (i.e. is not pure) after deconvolution in one of the data sets have a greater chance to be successfully discriminated.

2.2. Peak detection

The applied peak detection method has been previously reported [34]. This method requires no a priori information and has been reported to work well with similar datasets to those used in the current investigation. The method basically combines two of the most commonly used approaches for peak detection by firstly matching the peaks with a reference peak shape and then selecting the peaks with a threshold based on the noise level. The necessary settings such as typical peak widths and noise level are automated. The algorithm functions in the chromatographic domain and is applied to each extracted ion chromatogram (XIC), where it is theoretically easier to discriminate peaks from noise. It utilises a Gaussian second derivative (GSD) digital filter that simultaneously enhances the analytical signals and removes the baseline [35]. The resulting data set is a baseline and noise free reconstruction of the original data set with similar intensities and peak widths, but with a slightly different shape in the chromatographic domain (The typical resulting peak shape will more resemble the upper 2/3 of a Gaussian peak due to the structure of the filter coefficients). The peak detection algorithm adapts linearly to expected changes in peak width over time (most evident for isocratic data). Raw data peaks deviating from the expectations will receive a peak width in the corresponding reconstruction more close to the expected value. In such circumstances the area will also be influenced to some extent. The reconstructed peaks will however be better suited for the forthcoming step where the peaks belonging to the same sample component will be bunched. Moreover, the algorithm can be optimized to either detect low signal-to-noise (S/N) peaks or to be more sensitive to the detection of overlapping peaks and is functional for both isocratic and gradient data.

2.3. Peak classification and bunching

During this step, each component chromatogram and corresponding spectrum is reconstructed by bunching relevant peaks with principal component analysis (PCA) in a method referred to as the principal component variable grouping (PCVG), which is well described originally in the work by Ivosev et al. [36].

To increase the accuracy and reduce the complexity, each peak detected and reconstructed data set, $\mathbf{D}_{\mathrm{pd}}$, was first divided into a number of sub-sets before being subjected to PCVG. Each subset was selected from $\mathbf{D}_{\mathrm{pd}}$ by first finding the start and end points of the chromatographic peak with the highest intensity in the data set and then group all the peaks with maximum intensity within this retention time interval into a peak cluster. A peak bunching step as outlined below is then performed on this subset and the result is stored in a new data set and is removed from $\mathbf{D}_{\mathrm{pd}}$ before the next peak cluster is formed. The sub-set formation is repeated until all the peaks have been processed. Dividing peaks into smaller clusters in this manner is effortless when using peak detected data and this reduces the calculation times and it is often easier to estimate the number of components in each cluster than in the full data set. In addition, the peak widths are theoretically about the same within each peak cluster which can otherwise have an impact on some of the critical eigenvalues from the PCA.

PCVG basically initially performs PCA on Pareto scaled data in order to determine the number of relevant principal components (PC) to use, which ultimately corresponds to the number of chemical components in the sample. The PCs are uncorrelated linear combinations of the original variables where the first explains as much of the variability of the data and each succeeding PC explains as much of the variability as possible in the remaining data. In the current case, the variables are each m/z ratio. Correlated variables are oriented in the same direction in the multidimensional

PC-space and are grouped by firstly finding the variable with the longest distance from the origin and then include all variables within a pre-set angle. The original sub-set chromatograms in $\mathbf{D}_{\mathrm{pd}}$ corresponding to the grouped variables in the PC loading space are summed and become the first preliminary component chromatogram. The previously grouped variables are removed from the PC-space before the procedure is repeated until the number of component chromatograms corresponds to the number of relevant principal components. All peaks are then assigned to the corresponding preliminary chromatograms depending on where the highest correlation within the elution time window is obtained. The resulting output is herein referred to as initial component chromatograms and spectra. The cut-off level for the number of relevant PCs and the size of the angle can be set arbitrarily, which both influences the results. From this method, each initial component chromatogram, $\mathbf{C}_i$, is the sum of all the reconstructed peaks from $\mathbf{D}_{\mathrm{pd}}$ which have been assumed to belong to the same component according to the above described bunching procedure, while each initial component spectrum, $\mathbf{S}_i$, is the sum of each reconstructed peak over the retention time range that belongs to the mentioned initial component. Each component spectrum and chromatogram contain only signals originating from the previously peak detected data set and thus does not contain the noise and background signals which are present in the original data. This means that the resulting spectra can be used directly in the comparison with other spectra without further considerations such as only comparing a pre-determined number of top intensity peaks [8,10], or the requirement to remove illogical peaks [5]. The PCVG based method cannot, however, separate a coeluted peak in the same mass channel into two components if it has not been properly differentiated during the peak detection step. An additional multivariate curve resolution (MCR) step, such as alternating least squares (ALS) [37] or iterative target transformation factor analysis (ITTFA) [38], could be used to refine the results further but a common problem associated with these kinds of MCR techniques is rotational ambiguity which means that several mathematical solutions to Eq. (2) are feasible and that the optimal solution might violate to the common theory of chromatography [39]. Methods for forcing the solution to behave in as physicochemically correct manner as possible have been developed, but it is difficult to monitor the successfulness of these modifications automatically. The current algorithm was instead developed to handle possible impurities in the initial components by utilising the fact that the components can elute differently during the different separation conditions and this is used to extract purer components from the initial bunched components at a later stage. Thus the PCVG settings were set so that the total number of initial components is under- rather than overestimated.

2.4. Component matching

When all the peaks in the data set have been bunched into initial component chromatograms with their corresponding spectra, the process is repeated for a secondary data set, with the same sample, but different separation conditions, so that $\mathbf{C}_1$ and $\mathbf{S}_1$ and $\mathbf{C}_2$ and $\mathbf{S}_2$ are obtained from $\mathbf{D}_{\mathrm{pd1}}$ and $\mathbf{D}_{\mathrm{pd2}}$ respectively. The next step is to then compare the spectra of each initial component in the primary data set with the spectra from each initial component in the secondary set. A similarity value is calculated between all the initial components in both data sets having at least one mass ion in common and the matches are ranked according to the index. Techniques for eliminating unlikely spectral matches have been reported [8], but such processing always includes a risk of erroneous discrimination and is less important with the present computational and storage power [17]. Comparing initial component spectra with the current method, however, can reduce the number of necessary calculations by some order of magnitude in comparison to comparing every spectrum in the unprocessed data, depending on the complexity of the sample and the current S/N levels. The similarity index used herein is dependent on the direction of the comparison (i.e. comparing one component spectra in $\mathbf{S}_1$ with one component in $\mathbf{S}_2$ is not the same as vice versa). If the top ranked candidates are the same initial components regardless of the direction compared, the mass ions common to both initial components are stored in new matrixes $\mathbf{S}^*_{1,i}$ and $\mathbf{S}^*_{2,i}$ with the corresponding concentration profiles, $\mathbf{C}^*_{1,i}$ and $\mathbf{C}^*_{2,i}$. These are the final components with chromatograms and spectra and are herein simply referred to as the components. No other potential matching initial component that has received less than the highest ranking position is stored nor are the mismatched top candidates before the next component is considered. In this manner, the algorithm does not require any manual arbitrary threshold for the obtained similarity index for determining which match can pass this criterion as for other methods [5,32]. When all initial components have been processed, the spectra that matched between the data sets (now stored in $\mathbf{S}^*$) are removed from $\mathbf{S}$ before the procedure is repeated. This enables a higher possibility to extract minor mixture components within the predominant effluent components in the subsequent iteration. This step resembles the method described by Atwater et al. where the subtraction of reference spectra to measured GC–MS spectra allowed an improved identification of minor components in the resulting residual spectrum [12] and was later implemented in the BPM method described by McLafferty and Stauffer [21]. The procedure continues until no top ranked candidates are matching after testing all remaining components. All matched spectra are ideally true and pure component matches and the unmatched residual is ideally a true component that can find a match in the next iteration step or a component that is false or not present in the other data set.

2.5. Similarity index calculation

The approach of using one or a combination of some sort of similarity measurement for matching spectra has been considered in several works such as determining the number of common ions [7,25], utilising probability theory [5,21], using the match angle, with or without modifications [13,15–18,26,28,32], the similar Pearson correlation coefficient [19,22,23,26], the absolute, Euclidean or other distances [7,9,11] or by some sort of other similarity measurements [6,8,13,14,27]. In the current investigation, three types of similarity measurements are calculated between the current target initial component in the primary data set and each initial component of the group of candidate initial components in the secondary data set. These are then weighed to a single index value. The index should preferably generate a high value when the same initial components, acquired during different separation conditions, are compared and a low value when comparing to any other initial component. Ideally, the index should be able to discriminate spectral dissimilarities between true matching components due to instrumental and random artefacts and components that are not the same but have similar spectra. Generally, however, this will require a more complex algorithm. The implemented measures of similarity in the current work only consider the part of the initial component spectra, S, common to both target, S'_{T}, and candidate, S'_{C}, so that $S'_{\mathrm{T}_i} \cap S'_{\mathrm{C}_i}$. A candidate is defined as a component with at least one common mass ion to the target.

The first similarity measurement included is the squared Pearson correlation coefficient, a in Eq. (3), where the relation between the spectra in the target, S_{T}, and in the current candidate, S_{C}, are evaluated so that the inner relation of all mass ions of the components should be similar in order to score a high value.

$$a = \left[\frac{\sum_i (S'_{\mathrm{T}_i} - \bar{S}'_{\mathrm{T}})(S'_{\mathrm{C}_i} - \bar{S}'_{\mathrm{C}})}{\sqrt{\sum_i (S'_{\mathrm{T}_i} - \bar{S}'_{\mathrm{T}})^2 \sum_i (S'_{\mathrm{C}_i} - \bar{S}'_{\mathrm{C}})^2}} \right]^2 \tag{3}$$

Since it is generally more difficult to estimate the correct intensity level at lower S/N levels, decent correlation of the higher spectral intensities should be given a higher weight and this can be achieved by calculating the un-centred correlation *b*, also named the spectral contact or contrast angle, match angle or dot product distance, according to Eq. (4). This is very similar to Eq. (3), apart from the fact that each spectrum is not centred around its mean. Reducing the impact from a lower abundance spectrum in other similarity measurements has previously been used, simply by removing them [5], whereas, for others, the impact of the larger peaks is reduced [11].

$$b = \sum_i \frac{S'_{\mathrm{T}_i}}{\sqrt{\sum_i S'^2_{\mathrm{T}_i}}} \frac{S'_{\mathrm{C}_i}}{\sqrt{\sum_i S'^2_{\mathrm{C}_i}}} \tag{4}$$

The last similarity value, *c*, is calculated by Eq. (5) with the goal of characterising the intensity differences between the components. Since the same amount of the sample mixture is injected in both experiments for this type of application, and identical experimental conditions are used with the exception of different columns, the areas of the chromatographic peaks are expected to be near constant. Coeluting peaks can, however, cause ion suppression [3].

$$c = 2 \sum_i \frac{S'_{\mathrm{T}_i} S'_{\mathrm{C}_i}}{\sum_i (S'_{\mathrm{T}_i})^2 + \sum_i (S'_{\mathrm{C}_i})^2} \tag{5}$$

Each of the three similarity values varies between 0 and 1 and are independent of the designation of S_{T} and S_{C} (i.e. gives equal value for the same pair of compared spectra). The three terms are combined to a single score value, named the similarity index, *I*. Since it is generally easier to score a high value of *a* and *b* when the number of mutual spectra indices is low, these are weighed by means of the explained variance in the target component, *v*, as shown in Eq. (6),

$$I = av + bv + c \tag{6}$$

where *v* is determined according to Eq. (7). Combining two similarity measurements for the inner relation of the spectra (*a* and *b*) to the index increases the impact of that property in contrast to the similarity of the areas (*c*) and is also indicated to provide better results in terms of correct matches in comparison to only using *a* or *b* alone for the investigated data sets. As only the variance of the target is considered, the reversed designations of the component spectra provide different results as I_{TC} does not give the same value as I_{CT}. This is made intentionally in order to utilise the combined separation power from both data sets as explained in Section 2.4.

$$v = \sum_i \frac{S'^2_{\mathrm{T}_i}}{\sqrt{\sum_i S^2_{\mathrm{T}_i}}} \tag{7}$$

With decently matching components, the score will be close to 3. Component matching with poor values of the similarities mentioned above will result in a score close to zero. If the highest similarity index is obtained between the target and a candidate and that particular candidate also obtains the highest similarity index with the target when switching primary and secondary data sets, the match is considered to be the best possible. The probability that the correct match is obtained is dependent on the similarity to the second best match [8,18]. Complementary methods for increasing the similarity of spectra has been reported by others such as weighing intensities from different mass numbers [17,28], but this was not applied in the current method.

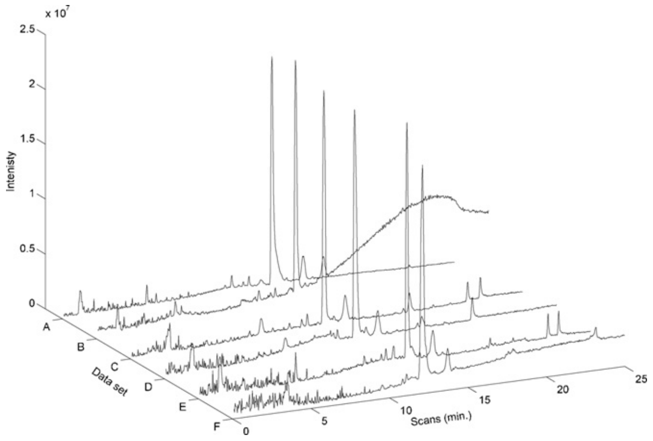

Fig. 2. TICs of the data sets, sampled with different columns. (A) Symmetry C18, (B) Altima HP C18 Amide, (C) Fluofix 120E, (D) Zorbax SB-CN, (E) ACE 3 phenyl, (F) Platinum EPS C18.

3. Experimental

3.1. LC–MS analysis

The proposed component tracking method was evaluated with experimental data sets which included typical variations in noise level, baseline drifts, signal distortions and other realistic artefacts that are usually present in such systems.

Six LC–MS data sets of one genuine sample of drug substance NN from AstraZeneca, spiked with 4% contaminants, was acquired using electrospray ionization and an Agilent Technologies 1100 Series MSD operated in the positive ion scan mode. A linear gradient from 5 to 95% ACN in water during 30 min was used. An acetic acid/ammonium hydrogen carbonate (3.9/10.0 mM) buffer giving a pH of 6.5, ion strength of 10 mM and a buffer capacity of 5.7 mM/pH unit was used. The injection volume was 5 μl. The columns used were 150 mm × 3 mm packed with ~3 μm particles: (A) Symmetry C18 (Waters Corp., Milford, MA), (B) Alltima HP C18 Amide (Alltech Assosiates Inc., Deerfield, IL), (C) Fluofix 120E (Thermo Hypersil-Keystone, Bellefonte, PA), (D) Zorbax SB-CN (Agilent Technologies Inc., Santa Clara, CA), (E) ACE 3 phenyl, Advanced Chromatography Technologies, Aberdeen, UK), (F) Platinum EPS C18 (Alltech Assosiates Inc., Deerfield, IL). The corresponding data sets are referred to as data set A–F and a TIC of each data set is shown in Fig. 2. Each data matrix consists of 5401 rows (*m*/*z*) and 703 columns (scans).

3.2. Data analysis

All calculations were performed using the MATLAB 7.0.4.365 (R14) Service Pack 2 (The MathWorks Inc., Natick, MA) on a PC with 3.0 GHz Intel Pentium 4 CPU and 4GB of RAM. The experimental data sets were converted from the Analyst LC–MS software (Applied Biosystems, Foster City, CA) by the wiff-to-matlab plugin.

3.3. Variable settings

The peak detection algorithm does not require any user input and was used with default settings as described in [34]. The PCVG method requires both a cut-off level for the number of significant components generated by PCA and the angle in the loading space for which the peaks in the sub-set are divided into concentration profiles and spectra. Both these settings have an impact on the outcome of the number of initially detected components and how well they are discriminated in the case of overlapping peaks. If these settings are set so that very poor resolving power is achieved for both data sets, there will be a greater risk that the final concentration

Table 1
Detected peaks and tracked components for the data sets.

Primary data set	Detected peaks	Initial components	Secondary data and number of tracked components					
			A	B	C	D	E	F
A	5267	297		263	164	199	238	271
B	4818	294	263		204	213	248	233
C	4541	328	164	204		235	201	231
D	6534	331	199	213	235		231	206
E	7622	332	238	248	201	231		214
F	6660	344	271	233	231	206	214	

profiles and spectra remain impure. In the present work, a static cut off level of 1% of the cumulative eigenvalues was used for the determination of the number of components. This includes a great deal of the total variation in each subset and proved to provide reasonable results since the majority of the original noise components is absent due to the peak detection step. The angle used was set to 45°, and this provided a reasonable trade-off between the resolution and the degree of allowed peak overlap.

4. Results and discussion

The results are presented with focus on the performance and accuracy of the algorithm when it comes to finding and tracking components when one data set is compared to another. Some results, however, are also given for the situation when several data sets are to be matched simultaneously.

4.1. Peak detection

Using the peak detection algorithm at maximum sensitivity to low S/N levels increases the risk of detecting false positives. Additional pre-processing can diminish this problem and could include a spike removal and/or an intensity threshold cut-off. The data sets were, however, used in an untouched manner in order to be able to investigate the performance deterioration caused by false positives and to be able to find and track even the smallest component in the sample by utilising the full performance of the instrument. The number of peaks found in the different data sets varied between about 4500 and 7600 as seen in Table 1. The large difference in the number of found peaks can be explained, in addition to the variation in the number of false positives, by the difference in the noise level in the data sets since only peaks passing a S/N level of 10 after signal enhancement are included by the algorithm. Moreover, all components in the sample mixture do not elute during the sampling period in some data sets so their presence is not recognised by the algorithm. An example can be seen in Fig. 2 where two late peaks in the TIC of data set C and E appear to be absent in data set A. Missing peaks due to too short sampling periods can and should naturally be avoided by for example sampling at a longer time and at a higher degree of organic modifier at the end of the runs. Missing peaks can also be the result of different adducts formation in the ion source due to different residues present in the columns used in spite of the fact that these had been thoroughly conditioned.

In Table 1, the number of tracked components is given, when using different data sets to be primary and secondary.

4.2. Peak classification and bunching

In the next step, peaks with similar retention times are assigned into initial components. The number of detected initial components in the data sets was 294–344 which can be seen in Table 1.

The initial components proposed by the algorithm for data set C were investigated manually. It was found, as expected, that many of the initial components actually were false positives due to detected spikes or low intensity noise signals surrounded by zero signals due to a minimum intensity threshold set by the instrument software during data acquisition. In such cases, the peak detection step is unable to estimate the noise level accurately and noise signals above the threshold can be regarded as peaks. The manual examination revealed that there were both many spikes in the data and many mass channels with portions of noise signal set to zero, generating both initial components consisting of only false positives but also initial components that partially consist of genuine peaks and partially of false positives. These artefacts can influence the performance of the algorithm. It was discovered that out of a total number of 328 initial components, as many as 177 were manually regarded as false, and these were mainly positioned at an early stage in the retention time. 35 initial components were regarded as being ambiguous and it was not possible to determine manually whether these were genuine or not due to low S/N levels together with few spectral indices. Thus, only 116 initial components (35%) were regarded as being assumed to contain true chemical information. These components did, however, include 83% of the estimated total area. Some of the automatically proposed initial components could be suspected of being impure, but this was not further investigated at this point. It is assumed that the other data sets have a similar distribution of false and informative components which complicates the following matching step since both the number of false targets and candidates increases.

4.3. Component matching

The spectral similarity index between each initial component in the primary data set is now calculated to the equivalents in the secondary data set and vice versa with the aim of finding the best matching components. As shown in Table 1, the number of tracked components varies between 164 and 271 depending on compared data sets and is identical between the same pair regardless of which one selected as primary and secondary. This means that the obtained results are insensitive to dataset assignation and no assumptions about the unprocessed or processed data sets are required. In Fig. 3, the 201 components that were automatically tracked between data set C and E are shown and the 25 components with the highest intensities are marked. These 25 matches were manually controlled and were correct except for the component numbers 12 and 14, which should be inter-mutually exchanged. These two components have an almost identical mass spectrum of near equal intensities and are thus difficult to distinguish. The smallest of the 25 marked peaks is number 5 with an estimated area of 0.3% of the main component.

The results obtained when using each of the data sets as the primary, and the subsequent data set in terms of alphabetic order as the secondary (data set F used A as secondary), were further investigated manually. The accuracy of the resulting component tracking together with the selectivity of the different columns can be viewed in Fig. 4, where the retention times of all the automatically matched components are plotted between the investigated data sets and manually classed depending on algorithm perfor-

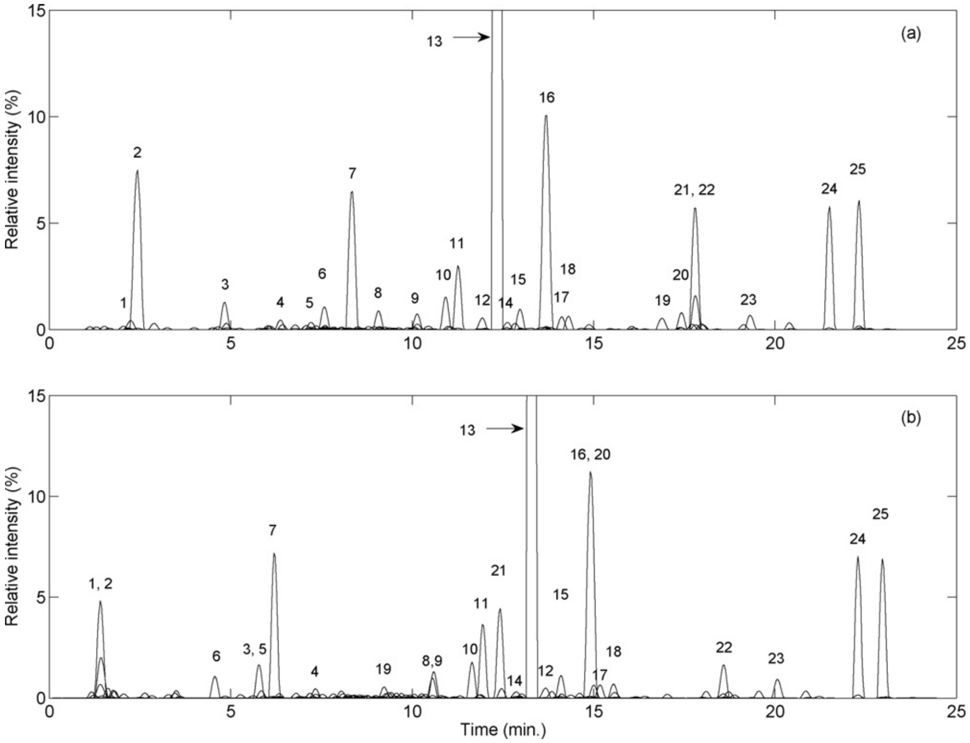

Fig. 3. Proposed component tracking between the primary data set C (a) and the secondary data set D (b). Numbers indicate the 25 most intense components.

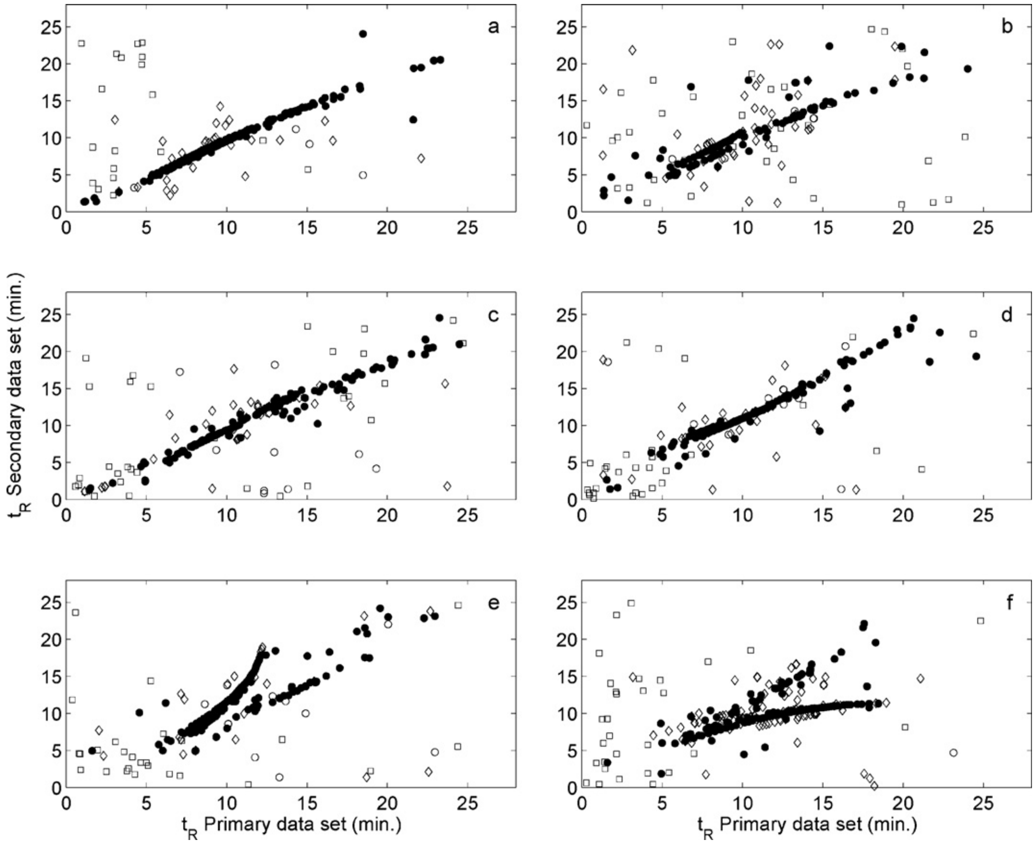

Fig. 4. Retention times of the automatically tracked sample components in two data sets. Each component is marked depending on the success rate of the algorithm according to manual inspection. (a) Data set A primary vs. data set B secondary, (b) B primary vs. C secondary, (c) C vs. D, (d) D vs. E, (e) E vs. F, (f) F vs. A. Markers: (●) correct match, (○) erroneous match between genuine components, (□) matched false positives, (◊) ambiguous matches. It should be noted that selectivity differences between the compared LC columns can lead to correct matches that substantially deviates from the diagonals in these plots.

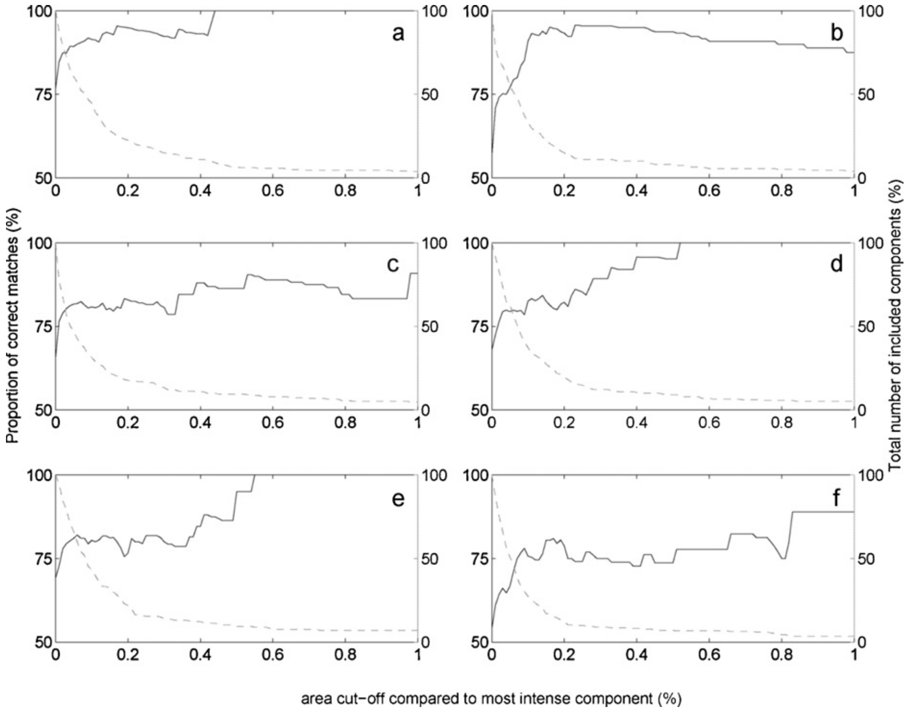

Fig. 5. The number of correct matches (-) and the total number of included components (–) at different minimum allowed area of the included components in the different primary data sets, (a) A primary, B secondary, (b) B primary, C secondary, (c) C primary, D secondary, (d) D primary, E secondary, (e), E primary, F secondary, (f) F primary, A secondary.

mance. It was concluded that most of the matches are correctly established automatically for these data sets, an average of 65% of the tracked components was clearly correct, whereas 3% was clearly an erroneous match between genuine components, and 13% was clearly not a genuine component (but still found a match in the secondary data set, genuine or not). The remaining components (19%) were regarded as ambiguous due to difficulties in manually establishing the correct match.

The main reasons for ambiguous matches are that some components contain equal spectra and are of similar heights making them difficult to differentiate, even manually, or difficult to validate as a genuine component due to low S/N levels. False component matches are often due to unexpected differences in intensity or the presence of false positives that surprisingly often had a counterpart present in both data sets. The majority of the false and ambiguous tracked components had, however, a small area and usually consisted of a single mass ion. Removing components based on single mass ions reduces the number of ambiguous and false positive matches by approximately 50% and the number of erroneous matches between genuine components by approximately 10%, whereas approximately 5% of the assumed correct matches are lost. Together, the single mass ion components had a total estimated area of 0.4% of the estimated total explained area on average and, as such, do not represent a substantial fraction of the informative parts of the data but did, however, significantly increase the number of correct matches.

In Fig. 5, the proportion of correct matches (according to manual assessment) together with the total number of components are shown for each data set used as a primary with the subsequent data set used as a secondary when disregarding tracked components below a certain area threshold. The area threshold is based on the area of the main component in the sample and is only shown between 0 and 1% in the figures. Since the number of considered components decreases monotonically at larger area cut-offs, a single erroneous match with high enough area can influence the proportion of correct matches negatively within the shown area interval (c.f. local maxima in Fig. 5b). The sample composition and the S/N levels in the data determine the minimum trackable area. At 0.05% level, the total number of correct matches reaches an average of 79% in the current data sets. It is evident that many of the tracked components have a small area and these are more difficult to match correctly. An average of 57% of the total number of tracked components had an estimated area above the 0.05% level. Combining a spike filter prior to peak detection with an area threshold before or after the peak bunching step could probably significantly diminish the problem associated with the erroneous matches. Another approach would be to increase the S/N cut-off level during peak detection. The default minimum S/N level allowed is 10, but with the enclosed S/N enhancement attribute of the peak detection algorithm used, the actual level can be even lower. At default settings in the peak detection step, peaks can theoretically obtain an increased S/N level of $0.55\sqrt{d}$, where d is the number of data points describing the chromatographic peak. In the current application, d is approximately 9, which means that peaks of an original S/N level of about 6 can be detected and, in theory, can be further processed. It is thus probable that a decrease in the sensitivity during peak detection could diminish the number of ambiguous components. Moreover, a minimum allowed similarity index value could be implemented during or after matching, but such a setting would be less intuitive for the analyst.

4.4. *Residual components*

The discriminated initial components containing the residual spectra should ideally not find a match during manual inspection. 319 of 328 of the initial components in data set C still contained one or more mass ions and it was possible to track 18 of these manually to a similar component in data set D. The main part of these components has lower S/N levels in one data set so that some or all of the ingoing peaks are never detected during the peak detection step. In some cases, the components contain artefacts (e.g. spikes) of higher intensities which obstruct the true match from

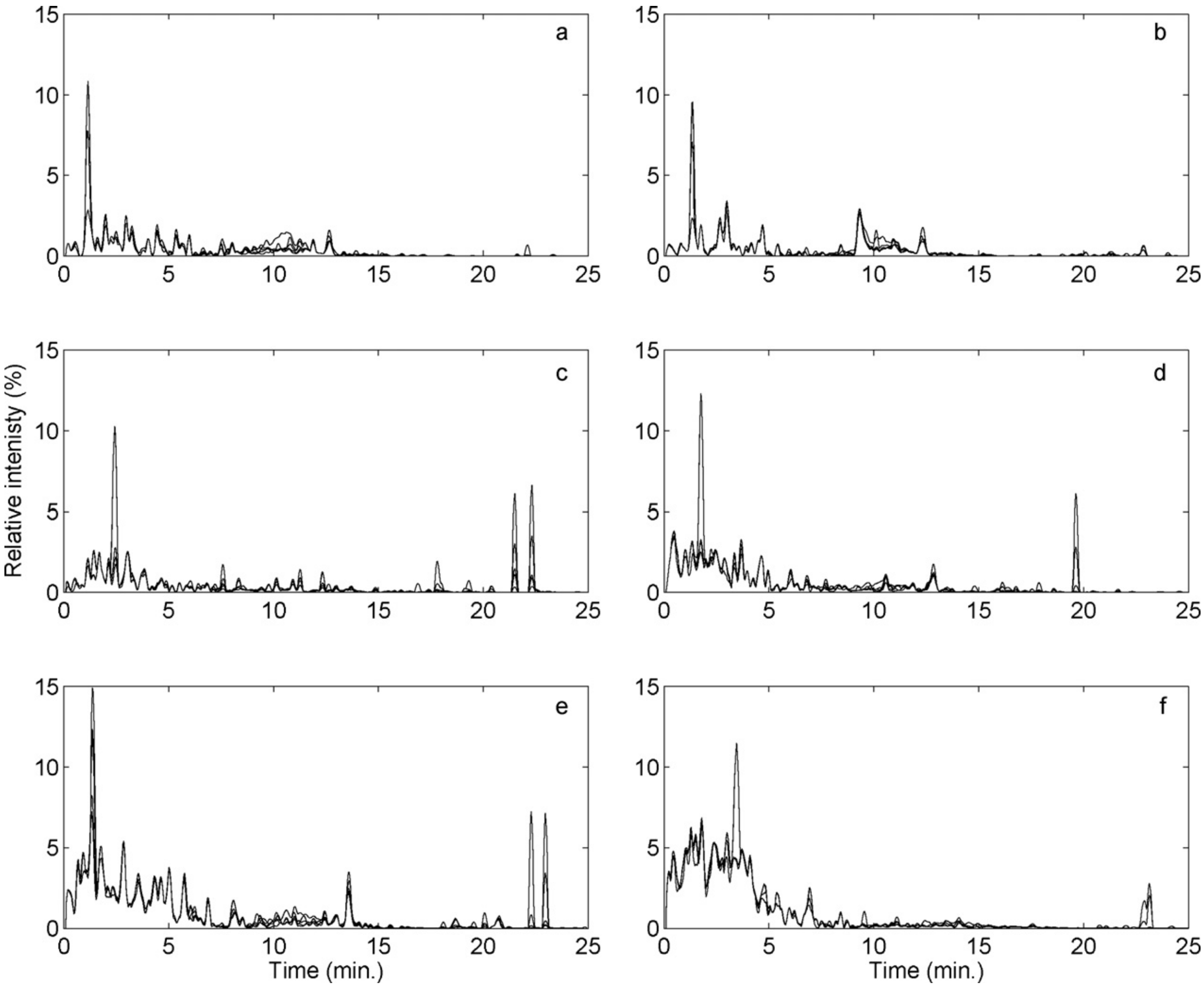

Fig. 6. Superimposed total ion chromatograms based on the residual spectra of primary data set (a) A, (b) B, (c) C, (d) D, (e) E, (f), F, after matching with all other data sets.

becoming the top candidate in one direction. It is often the case that these are positioned just under the top candidate in the list of matching candidates. The total estimated area of these 18 components (excluding artefacts) was 1.4% of the total explained area. It is possible that by removing the spikes prior to peak detection, the number of missed matching components can probably be significantly reduced. Chromatograms based on the residual spectra should not change considerably when matching between different data sets except at positions where true peaks are absent due to, for example, different adduct formations or because a sampling time range is too short. It is common for polar compounds with a low molecular weight to elute early and to generate signals of lower quality due to salts and residues that influence the electrospray and induce ion suppression. Fig. 6 shows that the chromatograms based on residual spectra are very similar regardless of the secondary data set used during the first minutes of the chromatograms. This indicates that these parts of the data sets are really non-informative and essentially lack chemical meaning since it cannot be matched regardless of separation condition used in the secondary data set.

4.5. Matching between several data sets

It is often of interest to match components between several data sets simultaneously. The proposed method can be used for matching components in a third or an arbitrary number of data sets by utilising the results from the previously matched data. The final matched spectra in $\mathbf{S}_1^*$ or $\mathbf{S}_2^*$ from data set one and two can be used as the initial components for the next data set, $\mathbf{S}_3$. In this way, only components present in $\mathbf{S}_1^*$ and $\mathbf{S}_2^*$ are searched for in $\mathbf{S}_3$. Then the final components in $\mathbf{S}_3^*$ are used as the initial components and are compared to $\mathbf{S}_4$ and so forth. When all the data sets have been processed, the components in the final data set, $\mathbf{S}_n^*$, are used as the initial components for the comparison with $\mathbf{S}_1$. The process can then be repeated once so that all data sets are processed with only those components present in all the considered data sets. By this approach, components that co-elute in several of the data sets can be successfully discriminated if they are decently separated in at least one data set. Constraints can be used to exclude irrelevant components or to increase the number of correct matches. In our case $n = 6$ and with the additional constraints that each component must have a minimum area of 0.05% and also a minimum number of two mass ions. This resulted in 66 components where the total number of clearly correct matches was 87.4% and the number of clearly erroneous matches was 2.3%. The remainder were regarded as being ambiguous. In Fig. 7, a truncated TIC of the data set A based on the components that could be tracked in all 6 data sets is visible. The movement in the retention time for a selection of correctly matching components in data set B–F is also shown. The minimum component area was estimated to $0.06 \pm 0.007\%$ (95% conf. level) of the main component. The estimated summed area of the correctly matched components present in all data sets was 79–92% of the total estimated component area. This indicates that the components of both small and large quantities could be tracked in the data sets.

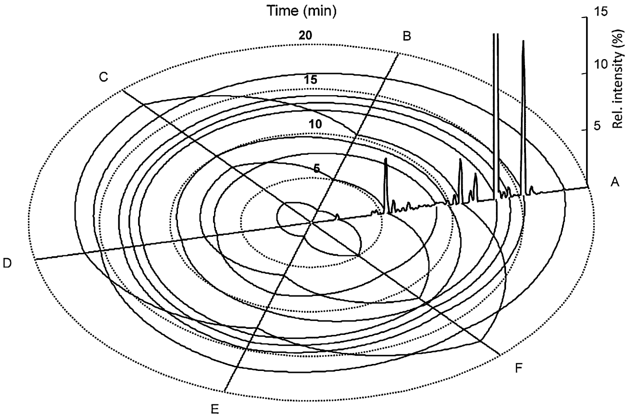

Fig. 7. A truncated TIC of data set A between 0 and 20 min and 0–14% relative intensity from the sample components present in all data sets and the retention time shifts of 8 selected components.

Some parts of the algorithm are computational intense and generally not fully optimized. The time required to perform a complete peak tracking is highly dependent on the number of detected peaks and components in the data. Further, the type of constraints used when several data sets are to be matched are of importance. For the six data sets used in this investigation, the peak detection step require approximately 70–90 s per data set, the peak bunching step approximately 10 s and the peak tracking step (between two data sets) required 90–150 s on a standard hardware equipped laptop computer. The time needed for tracking components between all six data sets (after peak detection and bunching) without the constraints used above was approximately 10 min. Enabling constraints typically reduces calculation times significantly.

5. Conclusions

Manual tracking of sample components acquired during different separation conditions is often very troublesome and time consuming since the elution order is highly unpredictable. The proposed method is capable of finding and tracking sample components automatically between two data sets, but can also be used to track common components between an arbitrary number of data sets. The confidence in the automatically tracked components can be increased by, for example, excluding proposed matches under a specified area or by setting the minimum number of allowed mass ions. Components not found in the second data set are excluded from the results. A list containing the best match together with all other candidates are presented so that the user can be alerted when candidates with a similar score to the top candidate are present. Components with a component area below the 0.05% level could be tracked in the current data sets, even in cases where some of the components are totally coeluted in one data set. In the current application, components in the sample were tracked when different columns were used, but the algorithm should also be functional when other separation conditions are to be tested, such as changing the temperature or the gradient profile.

Acknowledgements

We would like to thank AstraZeneca, Analytical Development, R&D Lund, Sweden, and the European Regional Development Fund of the European Union for their financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2010.10.083.

References

[1] S. Gorog, Trends Anal. Chem. 25 (2006) 755.
[2] M.R. Euerby, P. Petersson, J. Chromatogr. A 994 (2003) 13.
[3] L. Burton, G. Ivosev, S. Tate, G. Impey, J. Wingate, R. Bonner, J. Chromatogr. B: Anal. Technol. Biomed. Life Sci. 871 (2008) 227.
[4] G.T. Rasmussen, T.L. Isenhour, J. Chem. Inf. Comput. Sci. 19 (1979) 179.
[5] G.M. Pesyna, R. Venkataraghavan, H.E. Dayringer, F.W. McLafferty, Anal. Chem. 48 (1976) 1362.
[6] K.S. Kwok, R. Venkataraghavan, F.W. McLafferty, J. Am. Chem. Soc. 95 (1973) 4185.
[7] B.A. Knock, I.C. Smith, D.E. Wright, R.G. Ridley, W. Kelly, Anal. Chem. 42 (1970) 1516.
[8] H.S. Hertz, R.A. Hites, K. Biemann, Anal. Chem. 43 (1971) 681.
[9] S.L. Grotch, Anal. Chem. 42 (1970) 1214.
[10] S.L. Grotch, Anal. Chem. 43 (1971) 1362.
[11] L.R. Crawford, J.D. Morrison, Anal. Chem. 40 (1968) 1464.
[12] B.L. Atwater, R. Venkataraghavan, F.W. McLafferty, Anal. Chem. 51 (1979) 1945.
[13] K.X. Wan, I. Vidavsky, M.L. Gross, J. Am. Soc. Mass Spectrom. 13 (2002) 85.
[14] C.S. Tong, K.C. Cheng, Chemometr. Intell. Lab. Syst. 49 (1999) 135.
[15] M.E. Swartz, P.R. Brown, Chirality 8 (1996) 67.
[16] S.E. Stein, J. Am. Soc. Mass Spectrom. 6 (1995) 644.
[17] S.E. Stein, D.R. Scott, J. Am. Soc. Mass Spectrom. 5 (1994) 859.
[18] S.E. Stein, J. Am. Soc. Mass Spectrom. 5 (1994) 316.
[19] M.J. Sniatynski, J.C. Rogalski, M.D. Hoffman, J. Kast, Anal. Chem. 78 (2006) 2600.
[20] K.G. Owens, Appl. Spectrosc. Rev. 27 (1992) 1.
[21] F.W. McLafferty, D.B. Stauffer, J. Chem. Inf. Comput. Sci. 25 (1985) 245.
[22] W. Li, C.Q. Hu, J. Chromatogr. A 1190 (2008) 141.
[23] B.Y. Li, Y. Hu, Y.Z. Liang, L.F. Huang, C.J. Xu, P.S. Xie, J. Sep. Sci. 27 (2004) 581.
[24] Y. Hu, Y.Z. Liang, B.Y. Li, X.N. Li, Y.P. Du, J. Agric. Food Chem. 52 (2004) 7771.
[25] D.W. Hill, T.M. Kertesz, D. Fontaine, R. Friedman, D.F. Grant, Anal. Chem. 80 (2008) 5574.
[26] F. Gong, B.T. Wang, F.T. Chau, Y.Z. Liang, Anal. Lett. 38 (2005) 2475.
[27] F. Gan, J.H. Yang, Y.Z. Liang, Anal. Sci. 17 (2001) 635.
[28] W. Demuth, M. Karlovits, K. Varmuza, Anal. Chim. Acta 516 (2004) 75.
[29] A. Bogomolov, M. McBrien, Anal. Chim. Acta 490 (2003) 41.
[30] P. van Zomeren, A. Hoogvorst, P. Coenegracht, G. de Jong, Analyst 129 (2004) 241.
[31] G. Xue, A.D. Bendick, R. Chen, S.S. Sekulic, J. Chromatogr. A 1050 (2004) 159.
[32] S.J. Dixon, R.G. Brereton, H.A. Soini, M.V. Novotny, D.J. Penn, J. Chemometr. 20 (2006) 325.
[33] 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, J. Chromatogr. A 1107 (2006) 273.
[34] M.J. Fredriksson, P. Petersson, B.O. Axelsson, D. Bylund, J. Sep. Sci. 32 (2009) 3906.
[35] R. Danielsson, D. Bylund, K.E. Markides, Anal. Chim. Acta 454 (2002) 167.
[36] G. Ivosev, L. Burton, R. Bonner, Anal. Chem. 80 (2008) 4933.
[37] R. Tauler, Chemometr. Intell. Lab. Syst. 30 (1995) 133.
[38] P.J. Gemperline, J. Chem. Inf. Comput. Sci. 24 (1984) 206.
[39] M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi, M. Maeder, J. Chemometr. 20 (2006) 302.