# Time Alignment Algorithms Based on Selected Mass Traces for Complex LC-MS Data

# Time Alignment Algorithms Based on Selected Mass Traces for Complex LC-MS Data

Christin Christin,[†] Huub C. J. Hoefsloot,[‡] Age K. Smilde,[‡] Frank Suits,[§] Rainer Bischoff,[†] and Peter L. Horvatovich[†,*]

*Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands, Biosystem Data Analysis, Swammerdam Institute for Life Science, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands, and BM T.J. Watson Research Centre, Yorktown Heights, New York 10598*

Received November 7, 2009

Time alignment of complex LC-MS data remains a challenge in proteomics and metabolomics studies. This work describes modifications of the Dynamic Time Warping (DTW) and the Parametric Time Warping (PTW) algorithms that improve the alignment quality for complex, highly variable LC-MS data sets. Regular DTW or PTW use one-dimensional profiles such as the Total Ion Chromatogram (TIC) or Base Peak Chromatogram (BPC) resulting in correct alignment if the signals have a relatively simple structure. However, when aligning the TICs of chromatograms from complex mixtures with large concentration variability such as serum or urine, both algorithms often lead to misalignment of peaks and thus incorrect comparisons in the subsequent statistical analysis. This is mainly due to the fact that compounds with different *m*/*z* values but similar retention times are not considered separately but confounded in the benefit function of the algorithms using only one-dimensional information. Thus, it is necessary to treat the information of different mass traces separately in the warping function to ensure that compounds having the same *m*/*z* value and retention time are aligned to each other. The Component Detection Algorithm (CODA) is widely used to calculate the quality of an LC-MS mass trace. By combining CODA with the warping algorithms of DTW or PTW (DTW-CODA or PTW-CODA), we include only high quality mass traces measured by CODA in the benefit function. Our results show that using several CODA selected high quality mass traces in DTW-CODA and PTW-CODA significantly improves the alignment quality of three different, highly complex LC-MS data sets. Moreover, DTW-CODA leads to better preservation of peak shape as compared to the original DTW-TIC algorithm, which often suffers from a substantial peak shape distortion. Our results show that combination of CODA selected mass traces with different time alignment algorithm is a general principle that provide accurate alignment for highly complex samples with large concentration variability.

**Keywords:** LC-MS data sets • data processing • dynamic time warping • parametric time warping • correlation optimized warping • proteomics • metabolomics

## 1. Introduction

Time alignment is a critical step in the data preprocessing of comparative studies based on LC-MS analyses, which are widely used in 'omics' experiments. Without an accurate time alignment, nonlinear retention time shifts across different chromatograms lead to incorrect peak matching and invalid subsequent statistical comparisons. This is especially the case for highly complex data sets with large concentration variation, where incorrect alignment may result in misinterpretation of comparative 'omics' experiments.[1−3] The importance of time alignment in comparative biomarker discovery studies is emphasized by the increasing number of published time alignment applications and review papers on the subject.[1,4−29] These time alignment methods differ in their benefit functions as the criterion to construct the warping function that transforms the original retention time to the corrected retention time of the sample chromatogram. Since the throughput of proteomics experiments is constantly increasing, continuous improvement of automated time alignment methods is needed to align accurately the large amount of generated complex LC-MS data.

In this work, we focus on the modification of two widely used time alignment algorithms: Dynamic Time Warping (DTW)[5] and Parametric Time Warping (PTW),[6] and compare the results to our earlier work with the Correlation Optimized Warping-Component Detection Algorithm (COW-CODA).[4] Comparisons of time alignment quality of these algorithms using TICs or BPCs have been performed earlier for COW and DTW,[18,25] COW

[*] To whom correspondence should be addressed. E-mail, p.l.horvatovich@rug.nl; tel, +31-50-363-3341; fax, +31-50-363-7582.

[†] University of Groningen.

[‡] University of Amsterdam.

[§] IBM T.J. Watson Research Centre.

10.1021/pr9010124

Published on Web 01/13/2010

and PTW,[26] or all of the three methods.[24] These comparisons were performed on simple data sets having a low number of compounds and low concentration variability. The studies concluded that generally COW improved peak alignment and resulted in close location of the corresponding peaks in different chromatograms, but in some cases, DTW resulted in tighter alignment compared to COW. However, DTW is prone to distortion of peak shape, while COW preserved well the peak shape after alignment. PTW was reported to be faster than DTW or COW but less precise in terms of time alignment.[24]

Several modifications have been introduced to these algorithms in order to improve the quality of time alignment and to adapt to different data sets. DTW has been adapted to align data sets derived from capillary electrophoretic,[24] gas chromatographic,[18,25,30] and near-infrared spectroscopic data sets.[20] Peak shape distortions were observed and identified as major disadvantages of the DTW algorithm. This led to further work in order to preserve peak shape.[10,25] Tomassi et al.[25] showed that changing the value of the slope constraint affected the extent of peak shape distortion after alignment. The optimum value was obtained empirically depending on the characteristics of the data (e.g., the initial retention time shift). Clifford et al. introduced a variable penalty for each nondiagonal move in the warping path.[10] Even though these modifications were able to retain the peak-shape after DTW, they were intended to work on one-dimensional profiles only.

We show that both DTW and PTW fail in aligning the same compounds across multiple LC-MS data sets from complex proteomics or metabolomics samples, where many compounds with high concentration variability elute at similar retention times. The reason for this failure is that neither TICs nor BPCs provide information about $m/z$ values for the benefit functions of DTW or PTW. Some approaches that take the mass spectrometric information into account have been published. Most of these approaches do not work on the full data but rather on peak lists that require prior peak picking using various algorithms.[19,21−23] The quality of time alignment using peak lists thus depends considerably on the quality of the peak detection algorithms.

Only a few approaches use single stage mass spectra and thus separate the intensity information for peaks of different masses eluting at the same retention time.[19,31,32] The first two methods either use the entire mass spectrum and a complex gap penalty functions to avoid a large number of consecutive nondiagonal steps in DTW,[19,31] or use score functions which involve all peaks in the mass spectra and try to calculate the score due to the pure signal by removing noise contribution obtained with random reordering of peaks within mass spectra and setting each score below 0.2 to 0.[31] A third method uses the 200 mass traces containing the highest peaks in the chromatograms.[32] The latter method provides insufficient description since it does not describe the form of the benefit function and the exact method used to combine the intensity information of different mass traces.

In the present studies, we are using a Component Detection Algorithm (CODA)[33] to select high-quality mass traces from a complete chromatogram prior to alignment. We subsequently separate the signals of the selected mass traces in the benefit function of DTW and PTW by summing up the differences between sample and reference chromatograms using each selected mass trace separately.

Combining CODA with DTW or PTW required fundamentally different mathematical approaches as compared to COW,[4] since the selection of high quality mass traces is performed prior to the alignment procedure, while in COW-CODA, mass trace selection is part of the warping procedure and different mass traces are selected for each COW segment. The alignment process of DTW and PTW, however, uses the same set of mass traces over the entire chromatographic time range and mass traces selection must be done prior to warping.

Performance of DTW-CODA and PTW-CODA was compared to each other as well as to COW-CODA and to the one-dimensional time alignment approaches (DTW-TIC, PTW-TIC and COW-TIC). The sum of overlapping peak areas was used as criterion to judge the time alignment quality on label-free single stage LC-MS data sets obtained during comparative profiling studies for biomarker discovery using trypsin-digested human serum and acid-precipitated urine samples.

## 2. Material and Methods

**2.1. Computational Methods.** Time alignment is driven by time concordance of common peaks (compounds) that are shared between reference and sample chromatograms. The success of such a procedure depends on the capacity of an algorithm to find as many common peaks as possible with high accuracy and to use only the information from these common peaks in the retention time shift correction procedure (benefit function). In general, a time alignment algorithm for LC-MS data must have the following properties: (1) assuring that peaks with similar retention times but different $m/z$ values are considered as separate features in the benefit function to avoid merging of different peaks signal having similar retention time but different $m/z$ values; (2) considering only data from peaks that are shared between the reference and sample chromatograms in the benefit function; (3) discarding data containing a high level of noise; (4) taking peaks, background and noise distribution in the LC-MS data set into account locally; and 5) assuming that there are no changes in elution order of analytes between different LC-MS chromatogram.[8,31]

Taking local peak distribution into consideration is more difficult using DTW or PTW than COW. This is because DTW performs retention time alignment data point-by-data point instead of the segment-wise procedure of COW. Selection of local high-quality mass traces, which are the same in the reference and sample chromatograms, but which could be different for each time point is not possible in DTW, because changing mass traces for different retention time data points will lead to discontinuity in the calculated minimal cumulative distance used in the benefit function of the algorithm. As is the case for DTW, it is not possible to use different mass traces at different retention times in PTW, since PTW computes the warping function using an iteration procedure. In each step, it calculates the quadratic distance of the two traces using the entire time range, and locally different mass traces would result in similar discontinuity as with DTW. For that reason, we have introduced a global mass trace selection procedure, to measure the quality of mass traces across the entire chromatogram based on the average local quality of the mass traces. We have further used these selected high quality mass traces for the entire retention time range in the warping procedure.

**2.1.1. Measuring the Average Quality of LC-MS Mass Traces.** This section describes how the local quality of a chromatogram is determined by measuring the quality of a mass trace and how high-quality mass traces are selected prior to the warping procedure. The quality of a chromatogram corresponds to the ratio between peak related information and noise. Three main types of noise in mass spectrometry data are spikes, chemical noise and electronic noise. Signals that correspond to a single data point, the so-called spikes, are generated at the ion source between the LC and MS and the MS ion optics interface.[34] The other two noise components are due to ionized, contaminating chemical compounds (chemical noise)[35] and to the electronic noise from the detector. The local average of the combined noise is the local background level of the chromatogram.

A mass trace with high background will have a high mean value; thus, the mean subtracted mass trace will be rather different from the original signal. Similarly, a mass trace with spikes will differ strongly from the smoothed version that is obtained by using a moving average across several data points, as spikes are usually single-point events. Combining the measures of similarity between the mean-subtracted version and the smoothed version using a moving average of the original mass traces gives a single similarity value that takes both the contamination with spikes and the chemical and electronic background noise into account to result in a so-called quality index, the Mass Chromatographic Quality (MCQ) after Windig et al.[33] High quality mass traces contain low noise levels and low spikes relative to the intensity of detected peaks. The CODA algorithm selects mass traces containing a large number of high intensity peaks by calculating the MCQ of single mass chromatograms over the entire time range. However, in complex LC-MS data, it is often the case that the quality of LC-MS signals varies with respect to retention time even within a single mass trace. For the DTW and PTW algorithm, it is preferable to select a mass chromatogram containing a large number of peaks more or less evenly distributed across the entire time range rather than mass traces with similar MCQ values but containing few peaks that are concentrated in a narrow retention time window. We have therefore modified CODA to take the local peak distribution into account giving preference to mass traces that contain an evenly distributed high number of intense peaks.

The quality of the local signal for each mass trace is measured by applying CODA to overlapping moving windows with a length of $a$ data points, where $a$ is an odd number so that integer $b$ satisfies $a = 2b + 1$. Chromatogram $\boldsymbol{C}$ has size $m \times t$ where $m$ corresponds to the index of mass traces and $t$ to the index of retention times. For each position $(i,j)$ in chromatogram $\boldsymbol{C}$, an MCQ value of $C(i,j - b...j + b)$ is calculated. This MCQ value is regarded as the quality of the signal at position $(i,j)$. This step produces a matrix $\mathbf{Q}$, with the same size as the respective chromatogram, containing MCQ values for each data point of the chromatogram based on the equation below:

$$\mathbf{Q}(i,j) = \begin{cases} CODA(C(i, j - b...j + b)) \\ CODA(C(i, 1...2b + 1)) & \text{for} \quad j \leq b \\ CODA(C(i, t - 2b...t)) & \text{for} \quad j > (t - b) \end{cases} \qquad (1)$$

The quality of a mass trace is the average of the local MCQ values obtained for the same mass trace. Each chromatogram has thus a corresponding vector with size equal to the number of mass traces $m$. This vector contains the average MCQ values for the respective mass traces, and is used as quality scores to select mass traces prior to DTW and PTW (DTW-CODA, PTW-CODA).

**2.1.2. Mass Trace Selection for Time Alignment.** The alignment of two chromatograms requires the selection of several mass traces based on their respective quality in sample and reference chromatograms. Suppose chromatograms $\boldsymbol{C}_R$ and $\boldsymbol{C}_S$ have, respectively, a mean MCQ vector $A_R$ and $A_S$, where index $R$ and $S$ refer to reference and sample chromatograms, respectively. The product $A_S \times A_R$ indicates the combined quality of mass traces in both chromatograms. Mass traces that result the highest product are then selected to be included in the warping function. In this paper, the product of average MCQ values obtained with moving windows will be referred to as the "Local Component Detection Algorithm", abbreviated as LCODA.

**2.1.3. Dynamic Time Warping Combined with LCODA-Selected Mass Traces (DTW-CODA).** The DTW algorithm using one-dimensional signals has been described previously in a number of publications.[1,5,10,18−20,24,25] This section describes an extension of DTW algorithm using selected high quality mass traces based on the LCODA procedure. The concept of this method is illustrated schematically in Figure 1.

Suppose two chromatograms $\boldsymbol{R}$ (reference chromatogram with number of time scans $L_R$) and $\boldsymbol{S}$ (sample chromatogram with number of time scans $L_S$) are to be aligned using a set of LCODA-selected mass traces $\boldsymbol{K}$, for each combination $i$ and $j$, where $|i - j| \leq c$ (constraint). The local distance $d(i,j)$ is calculated by:

$$d(i,j) = \sum^{k \in K} (S_k(i) - R_k(j))^2 \qquad (2)$$

$\boldsymbol{S_k}(i)$ is the intensity value of mass trace $k$ at time point $i$ in chromatogram $\boldsymbol{S}$ and $\boldsymbol{R_k}(j)$ is the intensity value of mass trace $k$ at time point $j$ in chromatogram $\boldsymbol{R}$. Two matrices of size $L_S \times L_R$, which correspond to the grid presented in Figure 1, are constructed. The first matrix (*score matrix*) contains the minimal cumulative distances between $\boldsymbol{S}$ and $\boldsymbol{R}$. The second matrix (*path matrix*) contains the index of the optimum warping position that gives the respective cumulative minimum distance in the *score matrix*. The search space defining the allowed retention time transitions between the sample and reference chromatograms is defined by constraint $c$, which limits the maximal deviation from the diagonal by $c$ number of points. For each position $(i, j)$ in the defined search space, the minimum cumulative distances $D(i,j)$ are obtained from one of the three allowed predecessors $\{(i - 1, j), (i - 1, j - 1), (i, j - 1)\}$ based on eq 3. If the lowest score is obtained from different predecessors and one of them is the diagonal, then the diagonal path will be chosen to be included in the warping path.

$$D(i,j) = \min \begin{Bmatrix} D(i - 1, j) + d(i,j) \\ D(i - 1, j - 1) + d(i,j) \\ D(i, j - 1) + d(i,j) \end{Bmatrix} \qquad (3)$$

The global minimal distance between chromatograms $\boldsymbol{S}$ and $\boldsymbol{R}$ is obtained by means of dynamic programming from the *score matrix* by calculating $D(i,j)$ from position (1,1) until $(L_S,L_R)$ within the search space defined by constraint $c$. The global optimal warping path is obtained from the *path matrix* by backtracking the points resulting in the minimal cumulative distance as the last step of the time alignment procedure.

**2.1.4. Parametric Time Warping Combined with LCODA Selected Mass Traces (PTW-CODA).** The PTW algorithm for aligning one-dimensional signals between a sample $s(t_i)$ and a reference chromatogram $r(t_i)$ was introduced by Eilers.[6] The algorithm optimizes the coefficient $a_d$ of the polynomial warping function $w(t_i)$ with $d$ degrees so that the aligned sample signal

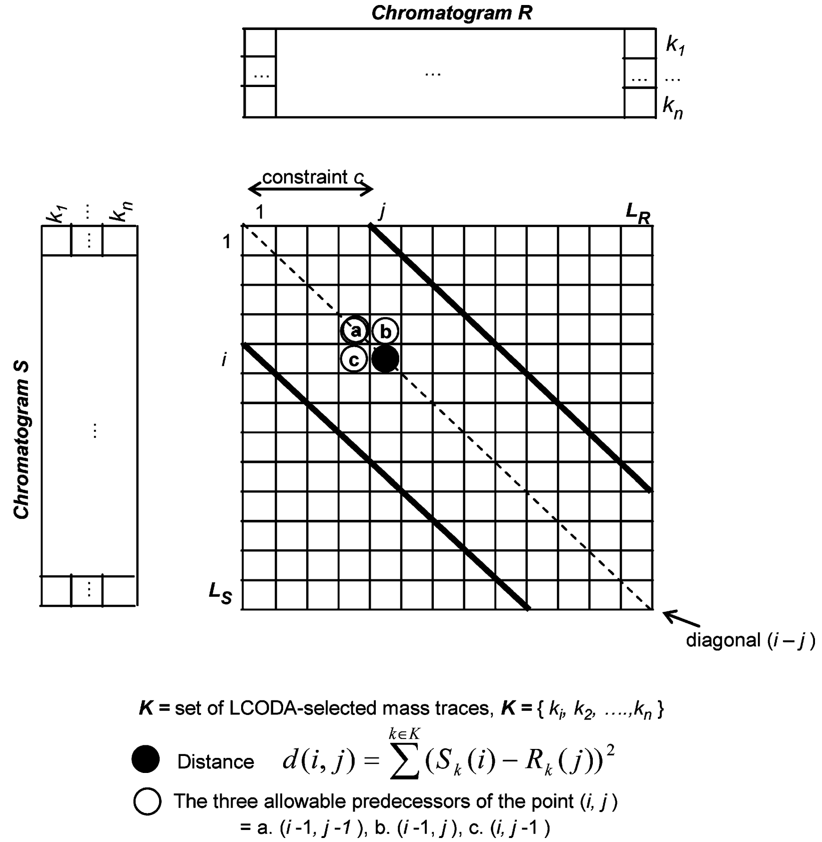
Chromatogram R
$k_1$
...
$k_n$
constraint c
1
j
$L_R$
1
i
$L_S$
a
b
c
diagonal (i – j)
Chromatogram S
$k_1$
$k_n$
K = set of LCODA-selected mass traces, K = { $k_i$, $k_2$, ….,$k_n$ }
● Distance $d(i,j)=\sum^{k\in K}(S_k(i)-R_k(j))^2$
○ The three allowable predecessors of the point (i, j) = a. (i -1, j -1 ), b. (i -1, j ), c. (i, j -1 )

**Figure 1.** Schematic representation of the DTW-CODA algorithm. The grid representation shows the dynamic programming approach to calculate the cumulative minimal distance between sample *S* and reference *R* chromatograms using a set of LCODA-selected mass traces *K*. The area rounded by bold diagonal lines represents the search space of the DTW algorithm as defined by constraint *c*. In this area, the cumulative minimum distance is calculated as the minimal sum of the intensities of LCODA-selected mass traces using the predecessor rules (see eq 3) starting from grid point (1,1) until reaching the final grid point ($L_S$, $L_R$). Score and path matrices corresponding to the grid coordinates contain the cumulative minimum distance and the grid location indices of the points according to the cumulated minimal distance obtained using predecessor rules and the constraint *c*. The final optimal warping path is determined by backtracking the preceding grid indices consecutively starting from ($L_S$, $L_R$) until reaching (1,1).

$s(w(t_i))$ has the lowest cumulative distance *G* to the reference. In the present study, we use a second-degree polynomial warping function with the form of $w(t_i) = a_0 + a_1t_i + a_1t_i^2$. The equation of the benefit function *G* is defined as follows:

$$G = \sum_{i\in H}[r(t_i) - \hat{s}(w(t_i))]^2 \qquad (4)$$

*H* indicates the set of indices *i* for which $\hat{s}(w(t_i))$ can be computed after interpolation of the sample chromatogram data points to the retention time vector (sampling points) of the reference chromatogram. To include the information of LCODA-selected traces, the benefit function *G* has been adapted to align two-dimensional signals based on a set of LCODA-selected traces ***K*** (eq 5). On the basis of a set of LCODA-selected traces, one obtains one warping function $w(t)$ using iterative process, that is used to calculate a newly aligned retention time vector for the sample chromatogram.

$$G = \sum_{k\in K}\sum_{i\in H}[r_k(t_i) - \hat{s}_k(w(t_i))]^2 \qquad (5)$$

**2.2. Property of the Data Sets and Data Preprocessing.** The algorithms were evaluated with three different LC-MS data sets with different analytical and biological characteristics as described in Christin et al.[4] Two data sets were derived from the analysis of trypsin-digested human serum (cervical cancer data set and factorial design data set) depleted of the six most abundant proteins and one data set from acid-precipitated urine of pregnant or nonpregnant women. The study protocol of the three data sets was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004. At the University Medical Center Groningen (UMCG, Groningen, The Netherlands) all newly referred patients are routinely asked to give written informed consent for collection and storage of pretreatment and follow-up serum, urine and tumor samples in a serum/urine/tissue bank for future research. Relevant patient data and follow-up are also retrieved and transferred into an anonymous, password-protected, database. Patients' identity is protected by study-specific, unique patient codes and their true identity are only known to two dedicated data managers. According to Dutch regulations, these precautions mean no further institutional review board approval is needed (http://www.federa.org).

All chromatograms were acquired on an ion trap mass spectrometer (Agilent Technologies, LC-MSD SL series, Santa Clara, CA) in randomized order with automated gain control of accumulation time of ions to reach fixed number of 30 000 ions in the trap, using a rolling average of two spectra in single-stage MS mode. The acquired data were converted and subsequently stored in centroid mode.

**2.2.1. Serum Samples.** Serum samples were obtained from the Department of Gynecological Oncology (UMCG) and stored at −80 °C in aliquots until analysis. Blood was collected in glass tubes (Becton Dickinson, Franklin Lakes, NJ) with a siliconized inner wall and allowed to clot for at least 2 h at room temperature before centrifugation at 1000 *g* for 10 min to obtain serum. Serum samples were stored at −80 °C in the local serum bank until use. Before LC-MS analysis, the serum samples were depleted of the six most abundant proteins using a Multiple Affinity Removal column (4.6 × 50 mm, Agilent Technologies). After trypsin digestion of the remaining proteins, all serum samples were stored at −80 °C in aliquots until the final LC-MS analysis. The depletion protocol and trypsin digestion is described in detail in Govorukhina et al.[36] Before the LC-MS analysis, 21 pmol of horse hearth Cytochrom C was added to 2 $\mu$L of serum samples. The most important tryptic fragments of horse heart Cytochrom C in the same type of serum, but different data set including retention time are listed in the bottom part of Table 1 in Horvatovich et al.[43]

**2.2.1.1. Cervical Cancer Data Set.** This data set is derived from a biomarker discovery study for cervical cancer. Serum samples from 10 patients were taken at two time points: before treatment (time point A) and after treatment with no recurrence of the disease for at least 6 months (time point B). These samples were analyzed by LC-MS resulting in 20 chromatograms. All patients in time point A showed high squamous cell carcinoma antigen-1 (SCCA-1) level (above 1.9 $\mu$g/mL) and no recurrence of disease after therapy except for two patients, one with partial remission and the other with stable disease where tumor remains without progression. The diagnosis was done by histological analysis and gynecological examination: inspection and palpation of the genitalia and SCCA-1 test. Patients with remission have no complains and normal SCCA-1 concentrations in time point B. All patients used in this study had advanced disease (stage III or IV) according to the International Federation of Gynecology and Obstetrics (FIGO) classification[37] and belonged to a group of long-term survivors. The level of the SCCA-1 was determined by ELISA.[38] Further details about the analysis of these samples using LC-MS are described in Govorukhina et al.[36]

**2.2.1.2. Factorial Design Data Set.** The serum sample for the factorial design study was obtained from one healthy female volunteer. The sample preparation procedure for this data set was similar to the cervical cancer data set except for the following seven factors that were varied deliberately to investigate the influence of preanalytical factors on the LC-MS profiles: blood collection tube, hemolysis level, clotting time, number of freeze−thaw cycles, trypsin to protein ratio, deactivation of trypsin after digestion, and stability of the digested sample in the autosampler of the LC-MS system at 4 °C. Each factor was varied at two levels (high and low), and from 128 possible combinations, 16 combinations were selected according to a two-level $2^{7-3}_{\mathrm{VI}}$ fractional design with resolution VI and with 3 repetitions of one condition. Detailed description of the factors and condition are described in Christin et al.[4] Nineteen LC-MS analyses were performed using the same protocol described by Govorukhina et al.[36]

**2.2.2. Acid-Precipitated Urine Data Set.** Twenty-five first-void midstream morning urine samples from pregnant women were obtained from a local biobank (Department of Obstetrics and Gynecology of the University Medical Center in Groningen, The Netherlands) and directly stored frozen at −20 °C. Twenty-five first-void midstream morning urine samples of nonpregnant women were collected in polypropylene containers and kept at 4 °C for a maximum of 1 day before the samples were stored at −20 °C in aliquots. The mass and retention time of spiked 7 individual peptides and the LC-MS method used to analyze the acid-precipitated urine samples are described in Kemperman et al.[39]

**2.3. Data Preprocessing.** The original LC-MS chromatograms were converted to ascii files using the Bruker DataAnalysis (version 3.4, Build 181) software. Each ascii file was transformed into a two-dimensional matrix containing intensity values using an in-house developed data preprocessing pipeline. In this matrix, each row has a corresponding $m/z$ and each column a respective retention time value. During transformation, data reduction from 0.1 to 1 amu per bin was performed in the $m/z$ dimension using two-dimensional Gaussian smoothing, while no data reduction was performed in the retention time dimension. All time alignment algorithms were applied to the transformed matrices. For each data set, one chromatogram was selected as the reference according to the procedure described in Christin et al.[4] Briefly, the best reference is the most frequently selected chromatogram having the highest sum of correlation to all other chromatograms in the data set based on the reconstructed TIC from a variable number of CODA selected mass traces. Similarly, the worst reference is the most frequently selected chromatogram having the lowest sum of correlation to all other chromatograms in the data set of the reconstructed TIC from a variable number of CODA-selected mass traces. The list of the best and the worst reference for each data set can be found in the Table S-1 (Supporting Information).

A peak picking algorithm was applied to the transformed matrices based on a filter developed by Radulovic et al. called $M - N$ rules[40] with $M = 8$ and $N = 3$. Signals are only retained if their intensity exceeds $n$ times the local baseline for $m$ consecutive data points in a single mass trace. The matrices obtained after peak picking are used later to evaluate and compare the quality of the time alignment algorithms by calculating the sum of overlapping peak area of two matrices. The data processing pipeline was executed on a personal computer equipped with an Intel Core Quad CPU Q9300 @ 2.5 GHz processor and 8 GB of RAM.

**Data Availability.** The complete three data sets used in this manuscript as test data will be available when the present article is accepted for publication at Tranche repository at https://proteomecommons.org/tranche/. Similarly, the code of DTW-CODA and PTW-CODA will be accessible when the publication is accepted at https://gforge.nbic.nl/.

## 3. Results and Discussion

**3.1. Importance of Data Preprocessing.** Ion trap mass spectrometers provide low resolution data, which contain small $m/z$ shifts of peaks caused by local space charge effects.[41] Binning procedures summing up intensity in mass spectra between predefined borders are often used to reduce the amount of data and thus the processing time.[19] However, binning procedures processing centroided ion trap data result in very noisy data because of the local space charge effect. Application of two-dimensional smoothing using a Gaussian kernel, on the other hand, results in smooth data in both dimensions, which contain less noise, especially in the retention time dimension, which leads to a lower accumulated error

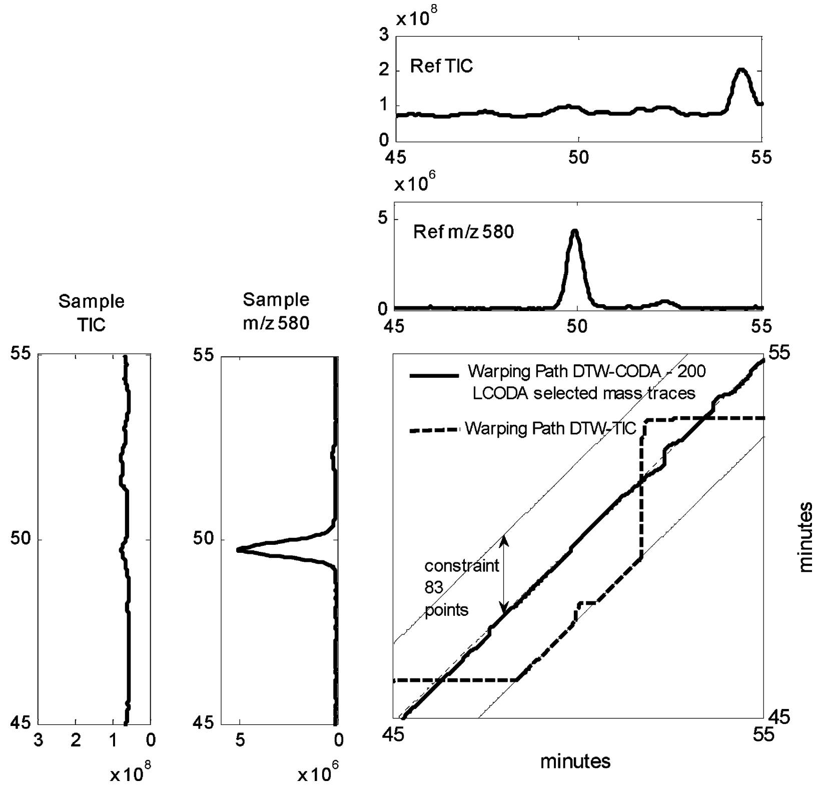

**Figure 2.** Comparison of the warping paths obtained with DTW-TIC and DTW-CODA using 200 high-quality LCODA-selected mass traces over a retention time window of 45−55 min in two chromatograms (5082628 and 5082630) from the acid-precipitated urine data set. DTW-TIC results in a rather 'chaotic' warping path ending in incorrect time alignment (Figure 3, middle panel). Using 200 preselected, high-quality mass traces, DTW-CODA follows a much smoother warping path correcting for the minor shifts in retention time that were present in the original data (Figure 3, top panel). Indeed DTW-CODA was able to find the optimal warping path by minimizing the sum of the cumulative distance of the LCODA-selected mass traces between the sample and reference chromatograms (see Figure 3, bottom panel).

in the benefit function of the time alignment algorithms (Figure S-1a,b). Figure S-1c shows an extracted ion chromatogram (EIC) of two adjacent masses of binned data and one mass trace of the data obtained after two-dimensional Gaussian smoothing of one peak. The binned data are fluctuating between the two adjacent mass traces, since the highest intensity is fluctuating between the borders of the bins. When mass spectra of binned data are used to calculate the correlation, such fluctuations between two adjacent mass traces will result in noise, as the fluctuation is a random event with respect to different chromatograms. On the other hand, data obtained by two-dimensional Gaussian smoothing will result in smooth Gaussian type profiles for each of the *m*/*z* mass traces providing thus an efficient contribution to the correlation between two chromatograms. In our application, we have used two-dimensional Gaussian smoothing to improve time alignment.

**3.2. DTW-CODA and PTW-CODA.** To compare the quality of the alignment with DTW and PTW in combination with CODA to the performance of the original algorithms (DTW-TIC and PTW-TIC, respectively), we applied them to label-free LC-MS data from complex biological samples. To present the operating principle of the DTW/PTW CODA algorithms, two chromatograms from the urine data set were chosen randomly. We selected chromatograms from this data set, because it is most challenging when it comes to time alignment problems due to the large concentration variation of compounds as a result of interindividual (biological) differences.

**3.2.1. Performance of DTW-CODA.** An internal standard peptide (YPFPG, *m*/*z* 580, retention time 49.3 min), which was not included in the 200 selected mass traces, was used to assess the local time alignment quality. A visual comparison of the alignment of this peptide by DTW-CODA using different numbers of LCODA-selected traces shows that a minimum of 20 selected mass traces is needed to reduce the extensive misalignments that were observed with DTW-TIC. However, peak shape distortions were only avoided when the number of selected mass traces was extended to more than 50. Selecting 200 mass traces proved to give the best alignment with negligible peak shape distortion.

The difference in the optimal warping path obtained with DTW-TIC and DTW-CODA indicates that the algorithms do not arrive at the same final warping function (Figure 2). This is due to the fact that the TIC of the sample chromatogram is rather different from that of the reference chromatogram in the depicted region (45−55 min) making it difficult, if not impossible, for the DTW warping function to find the optimal warping path, since mass spectrometric information is not considered separately. The result is a 'random' warping path that leads to poor alignment and a distorted peak shape (Figure 3, middle panel). Simplifying the initial alignment problem by selecting 200 high-quality mass traces allows the DTW algorithm to find a warping path that deviates little from the diagonal of the *score* and *path matrices* reflecting the true small shifts between retention times in the original sample and reference chromatograms (see Figure 3, top panel). Such a "smoother" warping path leads to time alignment without peak distortion and tight alignment of peaks (see Figure 3, bottom panel). On the other hand, DTW-TIC will only succeed in aligning LC-MS data sets

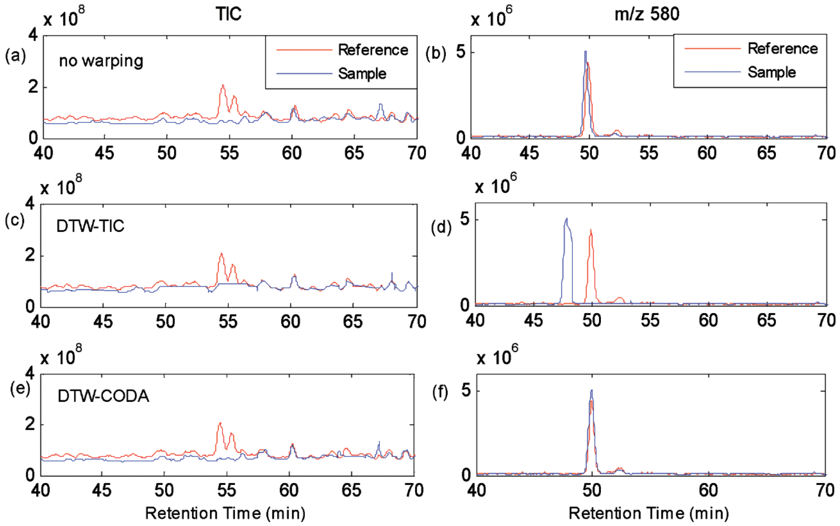

**Figure 3.** Application of DTW-TIC or DTW-CODA to two chromatograms obtained from acid precipitated urine sample (chromatograms 5082628 and 5082630; see also Figure 2. for the corresponding warping paths). (a) Original TICs (sample, blue; reference, red) prior to time alignment showing rather dissimilar profiles due to biological variability between samples; (b) original EICs (580 ± 0.5 amu) of the internal standard peptide YPFPG (*m/z* 580) prior to time alignment; (c) TICs after time alignment with the DTW-TIC algorithm showing peak distortions at various locations; (d) EICs (580 ± 0.5 amu) of the internal standard peptide YPFPG after time alignment with the DTW-TIC algorithm showing major distortion of the peptide peak and a larger retention time shift compared to the original data; (e) TICs after time alignment with the DTW-CODA algorithm showing tight alignment of the common peaks between sample and reference chromatograms even though the two most abundant peaks are absent in the sample chromatogram; (f) EICs (580 ± 0.5 amu) of the internal standard peptide YPFPG after time alignment with the DTW-CODA algorithm showing tightly aligned peaks without observable peak distortion.

when the TIC profiles are "well-defined" and similar as a result of little concentration variability in the analyzed samples. For that reason, time alignment using the TIC of complex 'omics' LC-MS analyses of body fluids containing many compounds with high concentration variability is challenging and in most cases impossible. Figure 3 shows the difference of the performance between DTW-TIC and DTW-CODA exemplified for a section of the EIC for *m/z* 580 related to the added standard peptide YPFPG. A major problem with the DTW-TIC algorithm is that it may lead to distortion of the chromatographic peak shape (see Figure 3, middle panel). When using several mass traces in the warping function of DTW-CODA, the algorithm tries to find the best compromise between alignments of each pair of peaks from different mass traces. This results in a smoother warping path and a decreasing number of consecutive nondiagonal moves resulting in less peak distortion.

Figure S-2 (Supporting Information) shows the overlaid two-dimensional image of the reference and sample chromatogram obtained from urine samples with the original retention time, and after correcting retention time shifts using DTW-CODA (present the entire elution range of compounds (42−92 min) over an *m/z* range between 80 and 620 amu). This figure shows that nonlinear retention time shifts in the sample chromatogram are accurately corrected with respect to the reference chromatogram by the DTW-CODA algorithm. A few peaks can be observed in the reference chromatogram (red) or sample chromatograms (green), which are absent in the corresponding other chromatogram (orphan peaks), while the shared common peaks present in both chromatograms (yellow) are well aligned. The retention times of these orphan peaks are also shifted by DTW-CODA to follow the same trend as the shared common peaks (e.g., peak at 335 *m/z* and 49 min of original retention time). This indicates that these peaks are also positioned correctly in the aligned chromatograms. In addition, the larger extent of retention time shifts observed in the beginning of the original chromatograms (20−40 min) (Figure S-2b), most probably due to the use of a trapping column, were effectively corrected. One major advantage of combining DTW with LCODA-selected mass traces is that, even without the use of special rules for the allowed predecessor steps in time alignment, the algorithm is highly conservative with respect to preserving the peak shape.

**3.2.2. Performance of PTW-CODA.** The same internal standard peptide (YPFPG, *m/z* 580, retention time 49.3 min) was used to assess the local alignment quality of PTW-CODA. To define the optimal number of selected mass traces, we calculated the sum of squared intensity differences using all points in single-stage LC-MS images between 8 randomly selected pairs of reference and the sample chromatograms after time alignment using a variable number of LCODA-selected mass traces ranging from 20 to 600. Two hundred selected mass traces resulted in the stabilized sum of squared intensity differences for these pairs of chromatograms (see Figure S-3 in Supporting Information). We have used the same 200 LCODA-selected mass traces for PTW-CODA as with DTW-CODA in order to facilitate comparison and to be sure that we are using a value, which is optimal for all chromatogram pairs.

The value of the 200 highest quality LCODA-selected mass traces gave accurate alignment in all three data sets. However, other data sets or different chromatogram pairs with different peak distribution or concentration variance may require a different number of high quality mass traces for optimal alignment. In that case, the above-described optimization procedure using the sum of squares of the intensity differences for all mass traces after applying the PTW-CODA algorithm should be used prior to the final application of PTW-CODA.

The main parameter to choose for the PTW algorithm is the degree of the polynomial of the warping function. It has been investigated that the alignment quality using a cubic warping function did not result in a significant difference to the

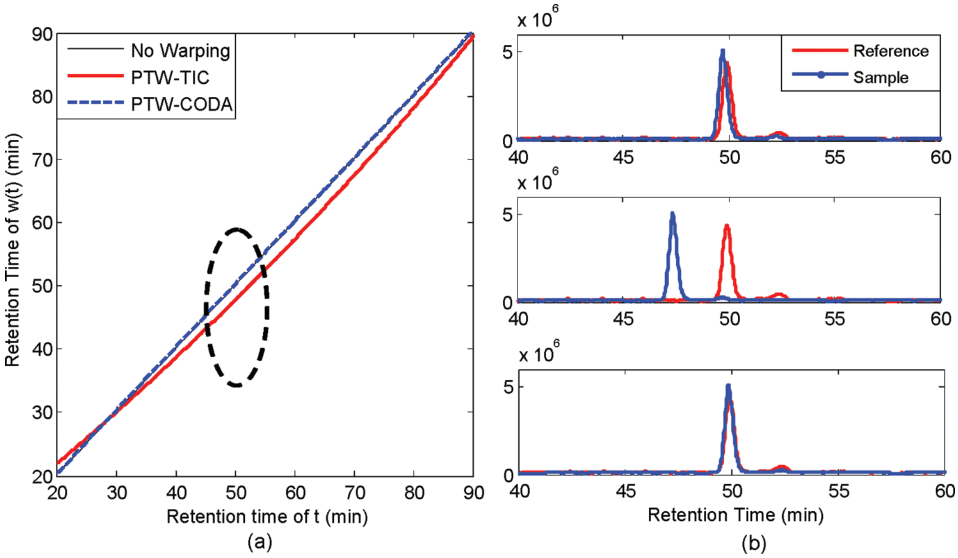

**Figure 4.** Comparison of the warping function of PTW-TIC and PTW-CODA using 200 high-quality LCODA-selected mass traces over a time window of 20−90 min applied to chromatograms 5082628 and 5082630 of the acid-precipitated urine data set. (a) Shows the warping function obtained by PTW-TIC (red line) and the warping function obtained by PTW-CODA (blue line). The ellipse indicates the region presenting almost the largest retention time shifts of spiked standard peaks between the reference and sample chromatograms after applying PTW-TIC. (b) The EICs of a standard spiked peptide YPFPG (*m*/*z* 580 and retention time 49,34 min) before alignment (top panel), after alignment with PTW-TIC (middle panel) and after alignment with PTW-CODA using 200 LCODA-selected traces (bottom panel).

alignment quality using a quadratic warping function, since the coefficient for the highest degree term was always close to zero (results not shown). This means that the quadratic function was able to accurately adjust to the true form of the nonlinear retention time shifts in the studied data sets. This may not be true for other data sets, where the true retention time shifts may have a more complex form. For that reason, if poor time alignment performance is observed with a quadratic function, using a higher order polynomial warping function may help to obtain a more accurate alignment. The only other user-defined parameter in PTW-CODA is the starting value of the coefficients of the warping function. As a starting point, we used the "no warping situation" in which the coefficients have the following values: $w(t) = t$, with $a_0 = 0$, $a_1 = 1$, and $a_2 = 0$.

Figure 4 shows the comparison between PTW-TIC and PTW-CODA using 200 LCODA-selected mass traces. The warping function of PTW-TIC deviates strongly from the diagonal in comparison with the warping function of PTW-CODA, which resulted in major misalignment of the corresponding peaks (see Figure 4b, middle panel). The largest retention time shifts between identical peaks were observed in the middle of the chromatograms, where the retention time shifts after warping were actually larger than in the raw data. The ellipse in Figure 4a indicates the region with the largest retention time shift after alignment with PTW-TIC. This resulted major misalignment of the standard peptide YPFPG (EIC $m/z$ 580 ± 0.5 amu) (Figure 4b, middle panel). In contrast, PTW-CODA was able to correct the slight retention time shift between the original chromatograms of the standard peptide (Figure 4b, bottom panel).

**3.3. Comparison of DTW, PTW and COW Coupled with LCODA-Selected Mass Traces.** We have shown that the combination of DTW, PTW, and COW with CODA or LCODA significantly improves the alignment quality compared to the original algorithms that use the TIC in the benefit function. The question remains which time alignment algorithm combined with CODA or LCODA will provide the best time alignment for a given experimental LC-MS data set. In this section, we compare the performance of DTW-CODA, PTW-CODA and COW-CODA by applying them to experimental data sets with different compound distributions in $m/z$ and retention time space (see Figure S-3 in Christin et al.[4]) and different concentration variability caused by contributions of various analytical and biological sources. The increasing order of the overall variance of the three data sets based on their respective relative standard deviation is: cervical cancer < factorial design < urine (see Figure 4c in Christin et al.).[4]

Concerning the COW-CODA algorithm, we have used the same values for the segment length and slack parameter as described previously.[4] Briefly, the segment length for the cervical cancer data set was 84 points (∼1.5 min), for the factorial design data set 139 points (∼2.3 min) and for the urine data set 83 points (∼2.2 min). The slack parameter was set to 20% of the given segment length. In the COW-CODA algorithm, the selection of high quality mass traces by CODA was performed segment-wise with a maximum number of 30 traces per segment. In cervical cancer, factorial design and urine data sets, the number of selected mass traces during the time alignment procedure (union of all traces of all segments) was 507, 396, and 307, respectively. The value of the global constraint $c$ in DTW was equal to the optimal segment length used in COW-CODA. Comparison of the performance of DTW-CODA and PTW-CODA was performed with 200 high-quality selected mass traces as discussed earlier. The constraint $c$ has to be chosen to satisfy $c > |L_R - L_S|$, which is generally not a problem even for ion trap mass spectrometry data, where the data dependent accumulation time results in small difference in the number of data points. However, if large differences in sampling rate occur, such as warping a chromatogram acquired in single-stage MS to a chromatogram acquired in MS/MS mode, the 3−5 times differences in sampling rate of single-

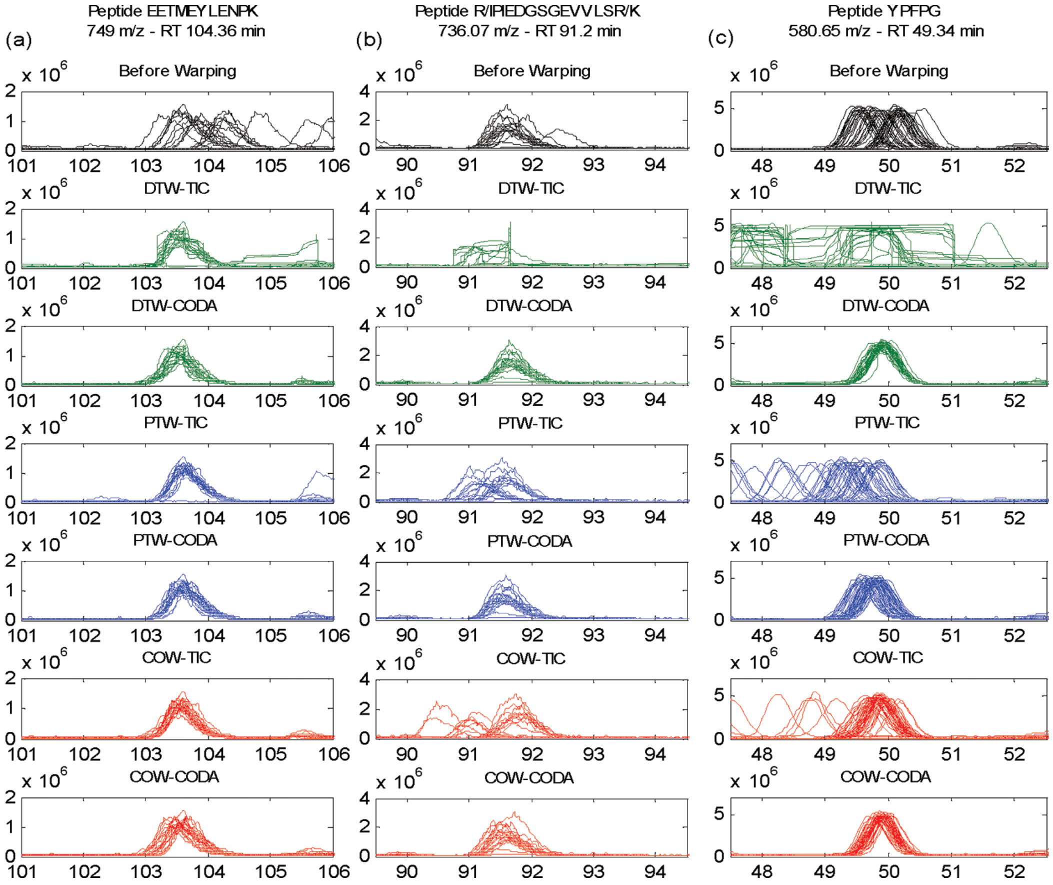

**Figure 5.** Local evaluation of the performance of the DTW-TIC and DTW-CODA (green) and the PTW-TIC and PTW-CODA (blue) algorithms in comparison with the previously described COW-TIC and COW-CODA (red) algorithms.[1] Extracted ion chromatograms show the retention time differences of spiked standard peptides in the cervical cancer (a), factorial design (b), and urine data sets (c) compared to the original data (top panel, black traces) using the best reference chromatograms.

stage MS information should be corrected by interpolation in order to apply the DTW-based algorithm with success.

The large dynamic concentration range of analytes and the accurate quantification to detect discriminating compounds between healthy and diseased states requires that LC-MS data is acquired in single stage MS mode for biomarker discovery. In that case, all time available for acquisition is used to collect quantitative information, while automated MS/MS would provide only quantitative information for every third or fifth scan time resulting in a decreased measured dynamic concentration range and less accurate quantification. After choosing the peaks of interest, the compound identity must be obtained from separate MS/MS measurements of pooled or individual samples.

To assess the performance of the algorithms, we chose data sets acquired in single stage MS mode. However, it is difficult to assess the true performance of time alignment algorithms using single stage MS data, since peaks cannot be related to identified compounds as compared to data obtained with automated MS/MS data acquisition, where peak identity is generally used to assess the accuracy of time alignment[19] or the overall performance of time alignment/peak matching algorithms.[42] In consequence, the alignment quality was evaluated locally by comparing EICs of added internal standard peptides that were present in each chromatogram of a given data set before and after time alignment with DTW-TIC, DTW-CODA, PTW-TIC, PTW-CODA, COW-TIC and COW-CODA using the best reference chromatograms. Visualization of the EICs of one standard peptide each in the cervical cancer (a), factorial design (b), and urine (c) data sets served to judge the time alignment accuracy of the different algorithms visually (Figure 5). In addition to the standard peptide peaks shown in Figure 5 and in the Supporting Information, we have visualized other standard peptides and numerous other peaks present in the majority of the samples. These peaks were distributed along the entire effective elution domain and showed similar time alignment characteristics to the 4 selected standard peaks presented in the current publication for each time alignment method. For all data sets, the original DTW-TIC algorithm showed the worst performance with considerable peak distortions and poor alignment while combining DTW with LCODA-selected mass traces (DTW-CODA) resolved these problems (see Figures S-4, S-5, S-6 in Supporting Information for other standard peptides). In general, all algorithms showed clearly improved alignment quality when combined with LCODA- or CODA-selected high-quality mass traces.

The initial concentration variability in the experimental LC-MS data sets plays an important role in the performance of the different time alignment algorithms. All algorithms (even DTW-TIC except for some remaining peak distortion) performed well on the cervical cancer data set (trypsin-digested serum; Figure 5a, left column) despite larger initial shifts in retention time, since the overall pattern was fairly conserved across the entire data set even at the TIC level. On the other hand, performance of time alignment methods was different for the factorial design data set (trypsin-digested serum; Figure

5b, middle column) containing larger analytical variability. For the factorial design data set, all of the TIC-based algorithms did not resolve retention time shifts for the standard peptides and even increased retention time shifts with respect to the original data. Combination of all alignment algorithms with CODA- or LCODA-selected mass traces improved the overall alignment quality and resulted in tight peak clusters. A similar tendency was observed for the more variable urine data set, where the algorithms that work with CODA- or LCODA-selected mass traces improve alignment quality or at least maintain the original quality of the data, in cases where retention time shifts were already low (see also Figures S-4, S-5, and S-6 in Supporting Information).

Our earlier observation showed that time alignment with COW-CODA is not sensitive to the choice of reference chromatogram. We confirm this important behavior for DTW and PTW algorithms combined with LCODA-selected mass traces (Figures S-4, S-5, S-6, for alignment with the best and Figures S-7, S-8, S-9 for the worst reference in the Supporting Information). It is thus not necessary to select the optimal reference chromatogram. It is, however, noteworthy that large nonlinear retention time shifts for peptides weakly binding to the chromatographic stationary phase, as sometimes observed at the beginning of the chromatographic elution gradient, are better corrected using the best reference chromatogram (see Figures S-6 and S-9 right column). The stability of the time alignment performance of algorithms using CODA- or LCODA-selected mass traces with respect to the reference chromatogram selection confirm further that all these methods results in tight alignment for the three experimental data sets.

To assess the global time alignment quality of different approaches, we compared the sum of the overlapping peak area between all possible pairs of chromatograms in the same data set as previously described.[4] An increased sum of the overlapping peak area is a measure of a globally improved time alignment. Figure 6 gives an overview of the sum of overlapping peak areas for each chromatogram with the remaining chromatograms in the three data sets using the best reference. The main observation is that all time alignment algorithms that make use of CODA- or LCODA-selected mass traces result in clearly increased overlapping peak areas when compared to the original data set independently of the variability in the original experimental data (Figure 6). The COW-TIC and the PTW-TIC algorithms show similar performance compared to the CODA-based algorithms for the cervical cancer data set, which contains the lowest compound concentration variability (Figure 6a). This result is in agreement with the local evaluation using EICs of spiked standard peptides and can be explained with the well-defined character and high similarity of TIC traces of the chromatograms in this data set. It is noteworthy that TIC-based algorithms may result in considerably higher retention time shifts than observed in the original data sets, especially for data set with high compound concentration variability. For example, in the factorial design data set, which has the lowest initial retention time shifts, all algorithms using TICs resulted in lower overlapping peak area than the original chromatograms for the majority of samples (see Figure 6b). This effect was less pronounced in case of the urine data set, where alignment quality was largely unaffected by the TIC-based algorithms (see Figure 6c). The global assessment of DTW-TIC is more difficult because of the large extent of peak distortion. The observed lower overlapping peak area for all three data sets shows that this method is not appropriate to align complex LC-MS data sets. In contrast to DTW-TIC, the DTW approach combined with LCODA-selected mass traces resulted in vast improvements and makes the DTW-CODA algorithm the most accurate in many instances.

The sum of overlapping peak areas obtained using the worst reference resulted in similar results as with the best references (Figure S-10). This confirms that the algorithms using high-quality CODA- or LCODA-selected mass traces perform well independently of the reference chromatograms. This omits the application of a reference selection method in the time alignment procedure.

## 4. Conclusion

Complex LC-MS data sets with many overlapping peaks in the retention time dimension are difficult to compare unless one can ensure that compounds are correctly matched prior to statistical analysis. We show that time alignment algorithms, that were originally developed for rather simple data sets, cannot be applied to this complex situation, as they either do not improve alignment or even make time alignment worse when using the TIC in a one-dimensional benefit function. We show furthermore that it is possible to simplify the initial complex data set by selecting $m/z$ traces based on their respective Mass Quality Indices (MCQ values) using a modification of the CODA algorithm originally described by Windig.[33] The combination of DTW, PTW or COW algorithms with CODA-based trace selection, considering the different selected mass traces separately in the benefit function, resulted in clear time alignment improvements in three different complex LC-MS data sets containing increasing levels of concentration variability. Local and global assessment of the performance of the new algorithms showed that they were successful in aligning complex data sets as obtained during biomarker discovery and other quantitative comparative proteomics or metabolomics studies. Furthermore, the time alignment algorithms using CODA- or LCODA-selected mass traces do not increase the number of parameters that need to be optimized. The optimal number of selected mass traces can be obtained from the data itself through optimization procedures.

A distinct advantage of DTW-CODA algorithm is that it does not use any gap penalty function next to the constraint $c$ to limit the search space. This facilitates the use of this algorithm, as compared to other versions of the DTW algorithm using separate mass information in their benefit functions.[19,31] The form of the benefit function, may also play a role. Future research should focus on better optimization of the form of the benefit function, such as using the cumulative covariance or the cumulative correlation.[19] We did not explore these possibilities, since we obtained highly accurate time alignments for all pairs of chromatograms in the studied data sets using the cumulative sum of quadratic distances as benefit function for DTW or PTW. We have shown that combination of CODA-selected mass traces with different time alignment methods is a general principle to align complex LC-MS data sets with high compound concentration variability. The main characteristics of the three time alignment algorithms based on our implementation in this work are presented in Table 1.

Although all three time alignment algorithms using mass trace selection perform similarly well on highly complex LC-

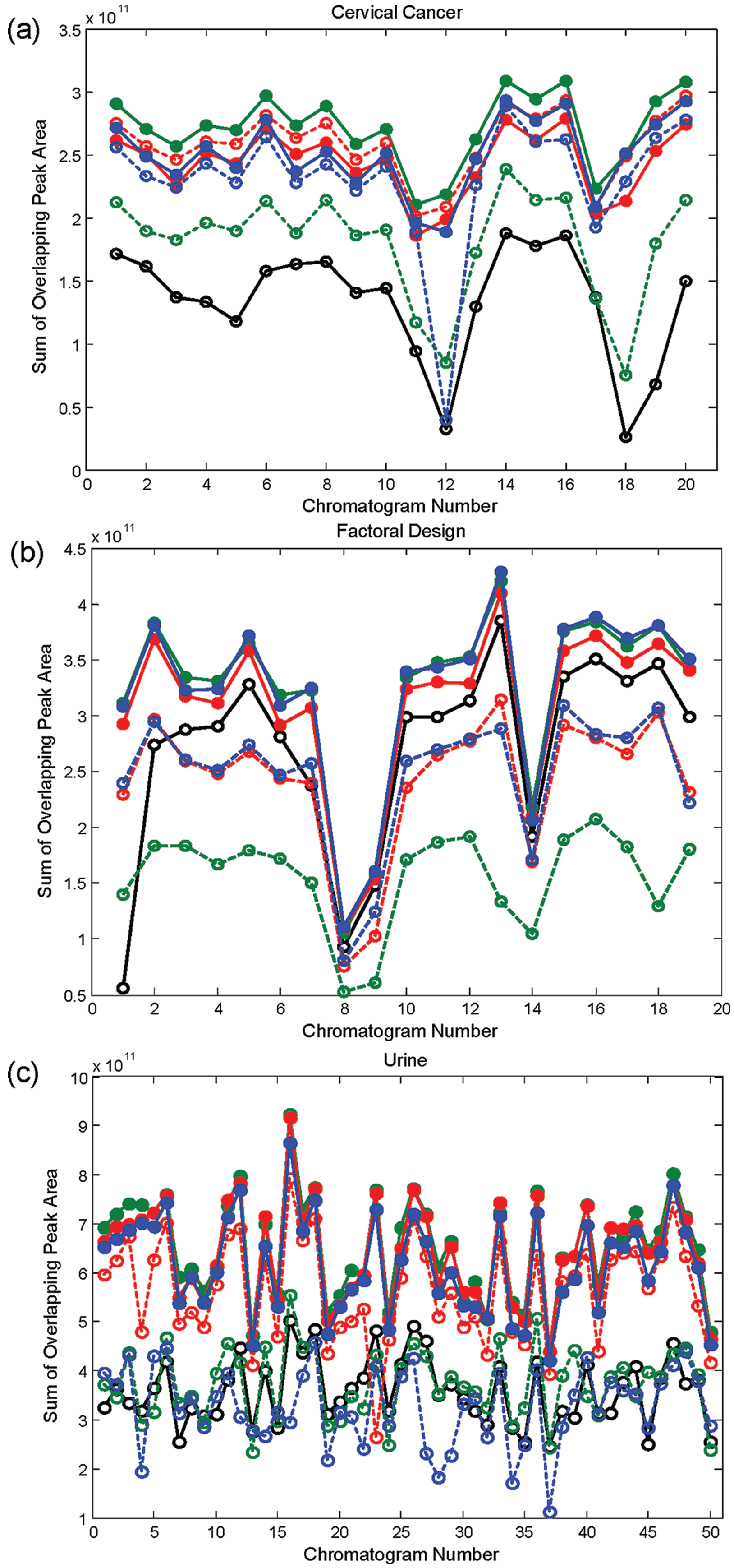

**Figure 6.** Sum of overlapping peak area of all chromatogram pairs using the best reference after applying $M - N$ rules as peak filter to the cervical cancer (a), factorial design (b), and urine (c) data sets. The original chromatograms before time alignment (black) are compared to the chromatograms obtained after alignment with COW (red), PTW (blue) and DTW (green) using TICs (dashed lines, empty circles) or CODA-/LCODA-selected mass traces (full lines, filled circles). The chromatogram names corresponding to the chromatogram indices in the figures are reported in Supporting Information (Table S2−S4).

**Table 1.** Main Characteristics of DTW-CODA, PTW-CODA and COW-CODA

| characteristics | DTW-CODA | PTW-CODA | COW-CODA |
|---|---|---|---|
| Setting of parameters | One parameter: constraint *c* (in time) to limit the borders of the search space. | No user-defined parameters required. | Two parameters: segment length (in time) and slack parameter (in % of segment length). |
| Peak shape distortion | Minor | No | No |
| CODA trace selection method | Global trace selection using the LCODA procedure. | Global trace selection using the LCODA procedure. | Local, segment-wise trace selection during algorithm execution. |
| Execution time for one pair of chromatograms[a] | ~15 s excluding mass trace selection (12 min per chromatogram). | ~1 min excluding mass trace selection (12 min per chromatogram). | 12 min including mass trace selection. |

[a] Total of 7000 time scans, 200 selected mass traces, using Intel Core Quad CPU Q9300 @ 2.5 GHz processor and 8 GB of RAM.

MS data sets, there are certain features, which discriminate them from a user perspective. The need to set parameters may complicate the proper use of an algorithm. User-defined values for parameters are an important point for some time alignment algorithms, since they affect the results significantly and should be adapted to the characteristics of the data sets (e.g., initial retention time shifts, average peak width, concentration variability). PTW-CODA has a distinct advantage in this respect for it does not require the user to set any parameters prior to starting the alignment procedure. The degree of the polynomial order of the warping function and the initial setup of the polynomial coefficients does not affect the alignment result, which is robust with respect to the changes of key analytical properties of the data sets such as peak distribution or concentration variation. Second, the requirements for computing capacity may be a limiting factor for users, especially if they do not have access to distributed computing facilities. DTW-CODA is advantageous in this respect, as it completes one round of time alignment for two complex LC-MS data sets in about 15 s on a powerful personal computer (Intel Core Quad CPU Q9300@2.5 GHz processor and 8 GB of RAM) while COW-CODA takes about 12 min. The significantly different execution time of COW-CODA is due to the fact that this algorithm includes segment-wise CODA trace selection as part of the warping function, while both DTW- and PTW-CODA require prior selection of high-quality mass traces, which takes 12 min per chromatogram. However, once the quality measurement by the LCODA procedure has been performed, this LCODA matrix can be reused for both DTW-CODA and PTW-CODA making this a one-time investment in computing time per data set. A note of caution has to be added with respect to using the DTW-CODA algorithm, as it may still introduce peak distortions, albeit much less than the original DTW-TIC algorithm. This must be carefully evaluated and can be considered as the main disadvantage of this approach. However, if peak quantification is performed before the time alignment and the alignment results are only used for peak matching, small peak distortions do not affect the statistical outcome of the comparative profiling study. In this case, time alignment will only affect peak clustering performance, which will be highly accurate as the chromatographic signal is tightly aligned in the data set, even though the peaks are slightly distorted.

To compare the performance of time alignment algorithms, the global evaluation based on overlapping peak areas is a reliable guide. However, since minor peak distortions and local misalignments may still occur, it is recommended to inspect the aligned chromatograms also visually using EICs of defined peaks (e.g., added internal standards or CODA-selected traces) before and after alignment.

**Acknowledgment.** The authors thank Natalia Govorukhina (serum analyses) and Ramses Kemperman (urine analyses) for providing the raw data as well as Ate van der Zee (University Medical Center Groningen: serum samples) and Frits Muskiet (University Medical Center Groningen: urine samples) for collaborations on cervical cancer biomarkers and urine analyses. We also thank Paul Eilers and Giorgio Tomasi for providing the Matlab code of PTW and COW to us, respectively. The work described in this publication was supported by grant BioRange 2.2.3 from The Netherlands Bioinformatics and The Netherlands Proteomics Centers.

**Supporting Information Available:** This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J. Chromatogr., A* **2002**, *961* (2), 237–244.

(2) Christensen, J. H.; Hansen, A. B.; Karlson, U.; Mortensen, J.; Andersen, O. Multivariate statistical methods for evaluating biodegradation of mineral oil. *J. Chromatogr., A* **2005**, *1090* (1−2), 133–145.

(3) Bahowick, T. J.; Synovec, R. E. Sequential chromatogram ratio technique: evaluation of the effects of retention time precision, adsorption isotherm linearity, and detector linearity on qualitative and quantitative analysis. *Anal. Chem.* **1992**, *64* (5), 489–496.

(4) Christin, C.; Smilde, A. K.; Hoefsloot, H. C. J.; Suits, F.; Bischoff, R.; Horvatovich, P. L. Optimized time alignment algorithm for LC-MS data: correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal. Chem.* **2008**, *80* (18), 7012–7021.

(5) Kassidas, A.; MacGregor, J. F.; Taylor, P. A. Synchronization of batch trajectories using dynamic time warping. *AIChe J.* **1998**, *44*, 864.

(6) Eilers, P. H. C. Parametric time warping. *Anal. Chem.* **2004**, *76* (2), 404–411.

(7) Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. r., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping. *J. Chromatogr., A* **1998**, *805*, 17–35.

(8) Aberg, K. M.; Alm, E.; Torgrip, R. J. The correspondence problem for metabonomics datasets. *Anal. Bioanal. Chem.* **2009**, *394* (1), 151–162.

(9) Chae, M.; Reis, R. J. S.; Thaden, J. J. An iterative block-shifting approach to retention time alignment that preserves the shape and area of gas chromatography-mass spectrometry peaks. *BMC Bioinf.* **2008**, *9*, S15.

(10) Clifford, D.; Stone, G.; Montoliu, I.; Rezzi, S.; Martin, F. o.-P.; Guy, P.; Bruce, S.; Kochhar, S. Alignment using variable penalty dynamic time warping. *Anal. Chem.* **2009**, *3*, 1000–1007.

(11) Finney, G. L.; Blackler, A. R.; Hoopmann, M. R.; Canterbury, J. D.; Wu, C. C.; MacCoss, M. J. Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution muLC-MS data. *Anal. Chem.* **2008**, *80* (4), 961–971.

(12) Fischer, B.; Grossmann, J.; Roth, V.; Gruissem, W.; Baginsky, S.; Buhmann, J. M. Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics* **2006**, *22* (14), 132–140.

(13) Fischer, B.; Roth, V.; Buhmann, J. M. Time-series alignment by non-negative multiple generalized canonical correlation analysis. *BMC Bioinf.* **2007**, *8*, 10>.

(14) Palmblad, M.; Mills, D. J.; Bindschedler, L. V.; Cramer, R. Chromatographic alignment of LC-MS and LC-MS/MS datasets by genetic algorithm feature extraction. *J. Am. Soc. Mass Spectrom.* **2007**, *18* (10), 1835–1843.

(15) Paulus, C.; Bonnet, S.; Gerfault, L.; Mery, E.; Strubel, G.; Ricoul, F.; Grangeat, P. Chromatographic alignment combined with chemometrics profile reconstruction approaches applied to LC-MS data. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2007**, *2007*, 5984–5987.

(16) Pierce, K. M.; Wood, L. F.; Wright, B. W.; Synovec, R. E. A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data. *Anal. Chem.* **2005**, *77* (23), 7735–7743.

(17) Pierce, K. M.; Wright, B. W.; Synovec, R. E. Unsupervised parameter optimization for automated retention time alignment of severely shifted gas chromatographic data using the piecewise alignment algorithm. *J. Chromatogr., A* **2007**, *1141* (1), 106–116.

(18) Pravdova, V.; Walczak, B.; Massart, D. L. A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta* **2002**, *456*, 77–92.

(19) Prince, J. T.; Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.* **2006**, *78* (17), 6140–6152.

(20) Ramaker, H.-J.; van Sprang, E. N. M.; Westerhuis, J. A.; Smilde, A. K. Dynamic time warping of spectroscopic BATCH data. *Anal. Chim. Acta* **2003**, *498*, 133–153.

(21) Sadygov, R. G.; Maroto, F. M.; Huhmer, A. F. ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. *Anal. Chem.* **2006**, *78* (24), 8207–8217.

(22) Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinf.* **2008**, *9*, 163.

(23) Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P. Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. *Anal. Chem.* **2008**, *80* (9), 3095–3104.

(24) Szymaska, E.; Markuszewski, M. J.; Capron, X.; van Nederkassel, A.-M.; Heyden, Y. V.; Markuszewski, M.; Krajka, K.; Kaliszan, R. Evaluation of different warping methods for the analysis of CE profiles of urinary nucleosides. *Electrophoresis* **2007**, *28* (16), 2861–2873.

(25) Tomasi, G.; van den Berg, F.; Andersson, C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemom.* **2004**, *18*, 231–241.

(26) van Nederkassel, A. M.; Daszykowski, M.; Eilers, P. H. C.; Heyden, Y. V. A comparison of three algorithms for chromatograms alignment. *J. Chromatogr., A* **2006**, *1118* (2), 199–210.

(27) van Nederkassel, A. M.; Xu, C. J.; Lancelin, P.; Sarraf, M.; Mackenzie, D. A.; Walton, N. J.; Bensaid, F.; Lees, M.; Martin, G. J.; Desmurs, J. R.; Massart, D. L.; Smeyers-Verbeke, J.; Heyden, Y. V. Chemometric treatment of vanillin fingerprint chromatograms. Effect of different signal alignments on principal component analysis plots. *J. Chromatogr., A* **2006**, *1120* (1−2), 291–298.

(28) Wang, P.; Tang, H.; Fitzgibbon, M. P.; McIntosh, M.; Coram, M.; Zhang, H.; Yi, E.; Aebersold, R. A statistical method for chromatographic alignment of LC-MS data. *Biostatistics* **2007**, *8* (2), 357–367.

(29) Zhang, D.; Huang, X.; Regnier, F. E.; Zhang, M. Two-dimensional correlation optimized warping algorithm for aligning GC x GC-MS data. *Anal. Chem.* **2008**, *80* (8), 2664–2671.

(30) Vial, J.; Nocairi, H.; Sassiat, P.; Mallipatu, S.; Cognon, G.; Thiebaut, D.; Teillet, B.; Rutledge, D. N. Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: application to plant extracts. *J. Chromatogr., A* **2009**, *1216* (14), 2866–2872.

(31) Prakash, A.; Mallick, P.; Whiteaker, J.; Zhang, H.; Paulovich, A.; Flory, M.; Lee, H.; Aebersold, R.; Schwikowski, B. Signal maps for mass spectrometry-based comparative proteomics. *Mol. Cell. Proteomics* **2006**, *5* (3), 423–432.

(32) Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **2003**, *75* (18), 4818–4826.

(33) Windig, W.; Phal, J. M.; Payne, A. W. A noise and background reduction method for component detection in liquid chromatography/mass spectrometry. *Anal. Chem.* **1996**, *68*, 3602–3606.

(34) Windig, W.; Smith, W. F.; Nichols, W. F. Fast interpretation of complex LC/MS data using chemometrics. *Anal. Chim. Acta* **2001**, *446* (1−2), 465–474.

(35) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **2008**, *627* (1), 71–81.

(36) Govorukhina, N. I.; Reijmers, T. H.; Nyangoma, S. O.; van der Zee, A. G. J.; Jansen, R. C.; Bischoff, R. Analysis of human serum by liquid chromatography-mass spectrometry: improved sample preparation and data analysis. *J. Chromatogr., A* **2006**, *1120* (1−2), 142–150.

(37) Benedet, J. L.; Bender, H.; Jones, H., 3rd; Ngan, H. Y.; Pecorelli, S. FIGO staging classifications and clinical practice guidelines in the management of gynecologic cancers. FIGO Committee on Gynecologic Oncology. *Int. J. Gynaecol. Obstet.* **2000**, *70* (2), 209–262.

(38) Esajas, M. D.; Duk, J. M.; de Bruijn, H. W.; Aalders, J. G.; Willemse, P. H.; Sluiter, W.; Pras, B.; ten Hoor, K.; Hollema, H.; van der Zee, A. G. Clinical value of routine serum squamous cell carcinoma antigen in follow-up of patients with early-stage cervical cancer. *J. Clin. Oncol.* **2001**, *19* (19), 3960–3966.

(39) Kemperman, R. F. J.; Horvatovich, P. L.; Hoekman, B.; Reijmers, T. H.; Muskiet, F. A. J.; Bischoff, R. Comparative urine analysis by liquid chromatography-mass spectrometry and multivariate statistics: method development, evaluation, and application to proteinuria. *J. Proteome Res.* **2007**, *6* (1), 194–206.

(40) Radulovic, D.; Jelveh, S.; Ryu, S.; Hamilton, T. G.; Foss, E.; Mao, Y.; Emili, A. Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2004**, *3* (10), 984–997.

(41) Cox, K. A.; Cleven, C. D.; Cooks, R. G. Mass shifts and local space charge effects observed in the quadrupole ion trap at higher resolution. *Int. J. Mass Spectrom Ion Processes* **1995**, *144* (1−2), 47–65.

(42) Lange, E.; Tautenhahn, R.; Neumann, S.; Gropl, C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinf.* **2008**, *9* (1), 375.

(43) Horvatovich, P.; Govorukhina, N. I.; Reijmers, T. H.; van der Zee, A. G.; Suits, F.; Bischoff, R. Chip-LC-MS for label-free profiling of human serum. *Electrophoresis* **2007**, *28* (23), 4493–505.

PR9010124