Optimized Time Alignment Algorithm for LC–MS Data: Correlation Optimized Warping Using Component Detection Algorithm-Selected Mass Chromatograms

ARTICLE *in* ANALYTICAL CHEMISTRY · SEPTEMBER 2008

Impact Factor: 5.64 · DOI: 10.1021/ac800920h · Source: PubMed

6 AUTHORS, INCLUDING:

Rainer Bischoff
University of Groningen
222 PUBLICATIONS **4,863** CITATIONS
SEE PROFILE

Peter Horvatovich
University of Groningen
61 PUBLICATIONS **884** CITATIONS
SEE PROFILE

Anal. Chem. **2008,** *80,* 7012–7021

Optimized Time Alignment Algorithm for LC−MS Data: Correlation Optimized Warping Using Component Detection Algorithm-Selected Mass Chromatograms

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

Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands

Correlation optimized warping (COW) based on the total ion current (TIC) is a widely used time alignment algorithm (COW-TIC). This approach works successfully on chromatograms containing few compounds and having a well-defined TIC. In this paper, we have combined COW with a component detection algorithm (CODA) to align LC−MS chromatograms containing thousands of biological compounds with overlapping chromatographic peaks, a situation where COW-TIC often fails. CODA is a variable selection procedure that selects mass chromatograms with low noise and low background (so-called "high-quality" mass chromatograms). High-quality mass chromatograms selected in each COW segment ensure that the same compounds (based on their mass and their retention time) are used in the two-dimensional benefit function of COW to obtain correct and optimal alignments (COW-CODA). The performance of the COW-CODA algorithm was evaluated on three types of complex data sets obtained from the LC−MS analysis of samples commonly used for biomarker discovery and compared to COW-TIC using a new global comparison method based on overlapping peak area: trypsin-digested serum obtained from cervical cancer patients, trypsin-digested serum from a single patient that was treated with varying preanalytical parameters (factorial design study), and urine from pregnant and nonpregnant women. While COW-CODA did result in minor misalignments in rare cases, it was clearly superior to the COW-TIC algorithm, especially when applied to highly variable chromatograms (factorial design, urine). The presented algorithm thus enables automatic time alignment and accurate peak matching of multiple LC−MS data sets obtained from complex body fluids that are often used for biomarker discovery.

Comparative proteomics and biomarker discovery studies often use label-free liquid chromatography coupled to mass spectrometry (LC−MS) to detect differences between preclassified sample sets. Easily accessible body fluids such as blood (plasma or serum) and urine are used primarily for this purpose. However, analyzing body fluids is challenging since they contain a large number of diverse compounds covering a wide dynamic concentration range leading to enormous amounts of raw data that need to be processed prior to statistical comparison. LC−MS data acquired in profile mode characterize compounds by their retention time and mass to charge ratio (m/z), and the quantity is reflected in the measured intensity (e.g., ion count).

Data processing workflows must be designed in a way to extract accurate information related to the identity and quantity of the detected compounds to allow subsequent statistical analyses and to find concentration differences between preclassified sample sets.[1] One of the most important challenges in detecting concentration differences is to ensure that identical peaks are compared across multiple samples (peak matching procedure), since liquid chromatography separations are prone to nonlinear elution time shifts as a result of slight variations in flow rate, gradient slope, and temperature as well as to column aging and the need to renew eluents from time to time. This is especially important for complex mixtures, such as depleted and trypsin-digested serum (shotgun proteomics approach[2]) or acid-precipitated urine, where many compounds elute with similar retention times. Improper correction of retention time shifts may thus lead to incorrect peak matching across multiple samples, resulting in statistical errors and the false discovery of biomarker candidates. Since clinical biomarker discovery and other proteomics or metabolomics applications require the comparative analysis of many samples to enhance statistical power, reliable, automatic, nonlinear time alignment algorithms are required to avoid such pitfalls.

Several techniques to correct nonlinear retention time shifts have been developed. These methods differ in both the search space and the benefit function (a measure of similarity) used to find the optimum retention time shift correction. Furthermore, most of the reported algorithms arbitrarily choose one chromato-

* 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.

[‡] Current address: Swammerdam Institute for Life Science, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.

[§] Current address: IBM T.J. Watson Research Center, Functional Genomics and Systems Biology Group, 1101 Kitchawan Rd., Route 134, Yorktown Heights, 10598 New York.

(1) Horvatovich, P.; Govorukhina, N.; Bischoff, R. *Analyst* **2006**, *131*, 1193–6.

(2) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999**, *17*, 676–82.

10.1021/ac800920h CCC: $40.75
Published on Web 08/21/2008

gram as reference and align all other "sample" chromatograms to it as a way to coalign all chromatograms.

The retention time vector in LC−MS data contains thousands of points. The search space, in which to find the optimal mapping between reference and sample retention time vectors, must be limited in order to keep computation time within reasonable limits and to avoid misalignment between distant, unrelated parts of the reference and sample chromatograms. Dynamic time warping[3−6] calculates the shift of data points in two chromatographic profiles and warps the trajectories in such a way that the distance between them is minimized using a set of constraints with respect to changes that are allowed for each point of the sample retention time vector. Correlation optimized warping (COW)[3,7−11] divides the chromatographic profile into segments and stretches or shrinks these in a linear manner within a limited search space to maximize the correlation to a reference chromatogram. Other approaches, such as parametric[12] and semiparametric warping,[9] use the full length of the chromatographic profile and perform the time correction in one step. Parametric warping optimizes polynomial coefficients by minimizing the difference of the intensity for a given data point between reference and sample chromatographic profiles.

The majority of the published time alignment methods use a one-dimensional benefit function[3,7,10,12−18] to search for the optimal alignment even when the data were acquired with a detector providing two-dimensional information in addition to the separation time (e.g., GC/MS, LC−DAD, LC−MS). For LC−MS data, the benefit function is often based on the total ion chromatogram (TIC, or sum of all intensities within one scan) or base peak chromatogram (or the maximal intensity in each scan). Time alignment using a one-dimensional benefit function may work well for samples where time and one-dimensional information used for alignment are similar across the samples, but it can be inappropriate for proteomics and metabolomics samples containing a high number of partially overlapping, closely eluting peaks with varying intensities. A few time alignment methods have been reported using a two-dimensional benefit function.[5,19−26] Some method0s specifically note the advantages of using a two-dimensional versus a one-dimensional benefit function.[5,24−26] Certain methods use single-scan mass spectra, known to be noisy due to scan-to-scan fluctuations,[23,27,28] whereas other algorithms use two-dimensional peaks that are local maximums of the ion intensity in the retention time and m/z space.[16,19,20,25,29−33]

Comparative studies have identified some of the advantages and disadvantages of warping algorithms using different search space and one-dimensional benefit functions on samples containing a small number of well-resolved compounds.[3,13,34,35] However, until now no evaluation and comparison of time alignment algorithms on complex data have been reported. Bylund[10] used covariance instead of the correlation coefficient as the benefit function to calculate the similarity between two chromatograms. This work concluded that the covariance measure is more sensitive to the peak height and will favor the alignment of large peaks, avoiding interference from regions containing mostly noise and background.

LC−MS chromatograms contain noise, background, and analyte peaks of varying quality with respect to the location in the chromatogram. Electrospray ionization, which is the most often used ionization technique for LC−MS of biomolecules, generates chemical noise and contaminants from solvents or the atmospheric environment that may be present at different parts of a chromatogram. Therefore, it is necessary to examine the local information content of LC−MS data and to locate regions containing high-quality information (low noise and background and a relatively high, compound-related signal) and to use those for time align-

(3) Tomasi, G.; van den Bergand, F.; Andersson, C. *J. Chemom.* **2004**, *18*, 231–41.
(4) Ramaker, H. J.; van Sprang, E. N. M.; Westerhuis, J. A.; Smilde, A. K. *Anal. Chim. Acta* **2003**, *498*, 133–53.
(5) Prince, J. T.; Marcotte, E. M. *Anal. Chem.* **2006**, *78*, 6140–52.
(6) Jaitly, N.; Monroe, M. E.; Petyuk, V. A.; Clauss, T. R.; Adkins, J. N.; Smith, R. D. *Anal. Chem.* **2006**, *78*, 7397–409.
(7) Nielsen, N. P. V.; Carstensen, J. M.; Smedsgaard, J. *J. Chromatogr., A* **1998**, *805*, 17–35.
(8) Fransson, M.; Folestad, S. *Chemom. Intell. Lab. Syst.* **2006**, *84*, 56–61.
(9) 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.; Vander, H. Y. *J. Chromatogr., A* **2006**, *1120*, 291–8.
(10) Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E. *J. Chromatogr., A* **2002**, *961*, 237–44.
(11) Sadygov, R. G.; Maroto, F. M.; Huhmer, A. F. *Anal. Chem.* **2006**, *78*, 8207–17.
(12) Eilers, P. H. *Anal. Chem.* **2004**, *76*, 404–11.
(13) Pravdova, V.; Walczak, B.; Massart, D. L. *Anal. Chim. Acta* **2002**, *456*, 77–92.
(14) Malmquist, G.; Danielsson, R. *J. Chromatogr., A* **1994**, *687*, 71–88.
(15) Athanassios, K.; John, F. M.; Paul, A. T. *AIChE J.* **1998**, *44*, 864–75.
(16) Johnson, K. J.; Wright, B. W.; Jarman, K. H.; Synovec, R. E. *J. Chromatogr., A* **2003**, *996*, 141–55.
(17) Listgarten, J.; Neal, R. M.; Roweis, T. S.; Emili, A. *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, 2005.
(18) Nordstrom, A.; O'Maille, G.; Qin, C.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 3289–95.
(19) Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M. *Bioinformatics* **2006**, *22*, 1902–9.
(20) Li, X. J.; Yi, E. C.; Kemp, C. J.; Zhang, H.; Aebersold, R. *Mol. Cell. Proteomics* **2005**, *4*, 1328–40.
(21) Prakash, A.; Mallick, P.; Whiteaker, J.; Zhang, H.; Paulovich, A.; Flory, M.; Lee, H.; Aebersold, R.; Schwikowski, B. *Mol. Cell. Proteomics* **2006**, *5*, 423–32.
(22) Radulovic, D.; Jelveh, S.; Ryu, S.; Hamilton, T. G.; Foss, E.; Mao, Y.; Emili, A. *Mol. Cell. Proteomics* **2004**, *3*, 984–97.
(23) Wang, P.; Tang, H.; Fitzgibbon, M. P.; McIntosh, M.; Coram, M.; Zhang, H.; Yi, E.; Aebersold, R. *Biostatistics* **2007**, *8*, 357–67.
(24) Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. *Anal. Chem.* **2003**, *75*, 4818–26.
(25) Fischer, B.; Grossmann, J.; Roth, V.; Gruissem, W.; Baginsky, S.; Buhmann, J. M. *Bioinformatics* **2006**, *22*, e132−40.
(26) Kirchner, M.; Saussen, M.; Steen, H.; Steen, J. A. J.; Hamprecht, F. A. *J. Stat. Software* **2007**, *18*, 1–12.
(27) Piening, B. D.; Wang, P.; Bangur, C. S.; Whiteaker, J.; Zhang, H.; Feng, L. C.; Keane, J. F.; Eng, J. K.; Tang, H.; Prakash, A.; McIntosh, M. W.; Paulovich, A. *J. Proteome Res.* **2006**, *5*, 1527–34.
(28) Finney, G. L.; Blackler, A. R.; Hoopmann, M. R.; Canterbury, J. D.; Wu, C. C.; Maccoss, M. J. *Anal. Chem.* **2008**, *80*, 961–71.
(29) Katajamaa, M.; Oresic, M. BMC *Bioinformatics* **2005**, *6*, 179.
(30) Johnson, K. J.; Prazen, B. J.; Young, D. C.; Synovec, R. E. *J. Sep. Sci.* **2004**, *27*, 410–6.
(31) Johnson, K. L.; Mason, C. J.; Muddiman, D. C.; Eckel, J. E. *Anal. Chem.* **2004**, *76*, 5097–103.
(32) Bergen, H. R., III; Vasmatzis, G.; Cliby, W. A.; Johnson, K. L.; Oberg, A. L.; Muddiman, D. C. *Dis. Markers* **2003**, *19*, 239–49.
(33) Palmblad, M.; Mills, D. J.; Bindschedler, L. V.; Cramer, R. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 1835–43.
(34) van Nederkassel, A. M.; Daszykowski, M.; Eilers, P. H.; Heyden, Y. V. *J. Chromatogr., A* **2006**, *1118*, 199–210.
(35) Szymanska, E.; Markuszewski, M. J.; Capron, X.; van Nederkassel, A. M.; Vander, H. Y.; Markuszewski, M.; Krajka, K.; Kaliszan, R. *Electrophoresis* **2007**, *28*, 2861–73.

ment. The component detection algorithm[36,37] (CODA) measures the information content of mass chromatograms containing a minimal amount of high-frequency noise, spikes (peak width of only one scan), and background by comparing the magnitude of change between the original trace and the mean-subtracted trace that was smoothed using a moving average. Hence, regions of high information content can be located and subsequently used for time alignment by COW.

In this paper, we combine mass chromatogram selection using CODA with a modified COW algorithm in order to take the local information in LC−MS chromatograms into account. The COW algorithm is applied segmentwise to pairs of selected mass chromatograms with the product correlation coefficient of the selected mass traces as a two-dimensional benefit function. The performance of the COW-CODA algorithm was evaluated using LC−MS data of urine and trypsin-digested human serum obtained from real-case proteomics and metabolomics studies.[38−40] These data sets exhibit different degrees of nonlinear retention time shifts and contain a large number of compounds of highly variable properties and amounts.

THEORY

Conditions for Proper Time Alignment Using COW-CODA. A number of conditions have to be met for successful application of the COW-CODA algorithm. Alignment is based on the presence of common peaks (compounds) between the reference and the sample chromatograms. The first criterion is thus to find a minimal number of common peaks in each time segment that should be aligned using the COW algorithm. In case there are no high-quality mass traces in a given segment, this segment is left unchanged, as there is no information to base the alignment on. This should, however, be the exception, if an overall well-aligned data set is to be obtained. The criterion of finding common, high-quality mass chromatograms is generally satisfied for all easily accessible body fluids in areas where compounds (peptides, metabolites) elute, since even highly variable body fluids such as urine contain a large number of conserved compounds.

We selected COW as the search algorithm for time alignment, because it corrects nonlinear time shifts by stretching or shrinking the data in a segmentwise fashion until optimal correlation has been reached. The limited search space of retention time range reduces the risk of large retention time corrections resulting in erroneous time alignments often occurring when the search space is to large. The aim of combining COW with CODA was 4-fold: (1) ensure that peaks with similar retention times but different m/z values are considered as separate features in the benefit function, (2) consider only data from common peaks between the reference and sample data sets in the benefit function, (3) avoid traces containing high noise and background, and (4) take into account that peaks, background, and noise distribution vary strongly among different regions of the LC−MS data set.

Component Detection by CODA. The CODA algorithm[36,37] was developed to select mass chromatograms with high-quality peaks, low noise, and low background. This algorithm contains two main steps: detection of spikes (single-scan signals originating from electronic noise) and detection of high signal background (typically originating from the mobile phase). When a mass chromatogram contains noise and spikes, the smoothed version will be different from the original chromatogram. We use a moving average to smooth the data in a specific m/z trace segment using a window larger than the peak width of the spikes, which results in large differences between the smoothed and original data when spikes are present and a low CODA similarity index. A mass chromatogram that has a high level of background noise will have a relatively high mean value. Hence, it will differ strongly from its mean-subtracted version resulting also in a low-similarity index. In contrast, a mass chromatogram with no spikes, a low level of background noise, and significant peaks will have a high similarity to its mean-subtracted version. The quality of a mass chromatogram is defined by a single, combined index of the two similarity indices. A high-similarity index thus indicates high-quality mass chromatograms with intense peaks. This similarity index is called mass chromatographic quality (MCQ) with a minimum value of 0 and a maximum value of 1. The CODA algorithm has been described in detail by Windig et al.[36,37]

Combining COW and CODA (COW-CODA). *Segmentation and Search Space.* The COW algorithm, as described by Nielsen et al.,[7] was employed in our procedure, with a modification of the benefit function, which was originally based on the sum of correlation of each segment using one-dimensional information (e.g., TIC or single-wavelength UV traces). Assume that we want to align a sample chromatogram S to a reference chromatogram R, with the number of scans in the reference and sample chromatograms being L^S and L^R, respectively. Each chromatogram may vary in length due to the different number of mass spectrometric scans (especially for ion trapping instruments). In our case, the number of mass traces d in each chromatogram is equal since we have used the same mass range for data acquisition (100−1500 Da) and both chromatograms have been identically smoothed from their original 0.1 amu resolution to 1 amu (see Gaussian smoothing and data reduction in the Materials and Methods). In the COW algorithm, the reference chromatogram remains unchanged while the end points of the sample chromatogram segments are allowed to move according to three constraints: (1) start and (2) end points of the sample chromatograms are unchanged, and (3) sample segments lengths are allowed to changed with the slack parameter; see point 3 of the Theory section in the Supporting Information (SI)). The flowchart of the warping process for two chromatograms is schematically described in Figure 1, while the detailed steps of the warping algorithm are described in the SI. COW specifies the degree to which the segment end points may move through the "slack" parameters. Details about the search space given by the slack parameters are discussed in Nielsen et al.[7] and in the Theory section of the SI.

Segmentwise Mass Chromatograms Selection. The procedure for selecting a high-quality mass chromatogram is the same for each segment. The MCQ values from the CODA algorithm were calculated for each mass chromatogram from a given segment of

(36) Windig, W.; Smith, W. F. *J. Chromatogr., A* **2007**, *1158*, 251–7.
(37) Windig, W.; Phalp, J. M.; Payne, A. W. *Anal. Chem.* **1996**, *68*, 3602–6.
(38) Horvatovich, P.; Govorukhina, N. I.; Reijmers, T. H.; van der Zee, A. G.; Suits, F.; Bischoff, R. *Electrophoresis* **2007**, *28*, 4493–505.
(39) Kemperman, R. F.; Horvatovich, P. L.; Hoekman, B.; Reijmers, T. H.; Muskiet, F. A.; Bischoff, R. *J. Proteome Res.* **2007**, *6*, 194–206.
(40) Govorukhina, N. I.; Reijmers, T. H.; Nyangoma, S. O.; van der Zee, A. G.; Jansen, R. C.; Bischoff, R. *J. Chromatogr., A* **2006**, *1120*, 142–50.

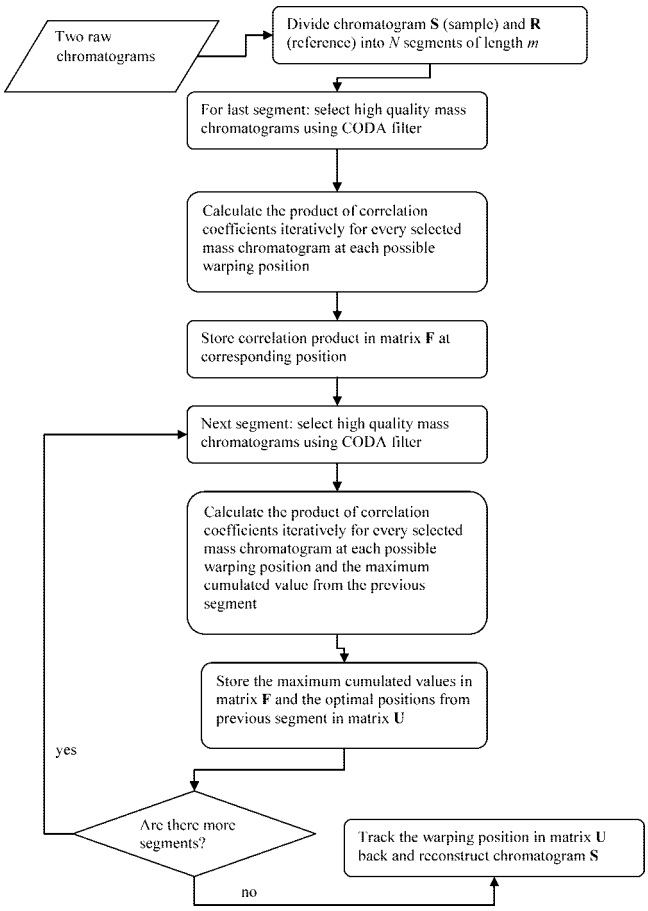

Figure 1. Flowchart of the COW-CODA algorithm, showing the main steps in warping a sample chromatogram S to the reference chromatogram R. This process is repeated until all sample chromatograms have been aligned to the chosen reference chromatogram in the data set.

chromatogram S and R, resulting in two vectors of length d containing the MCQ values of the corresponding traces. Since the segment end points of the reference chromatograms are fixed and the segments end points of the sample chromatograms can vary, a larger segment is chosen for the sample chromatogram compared to the corresponding segment in the reference chromatogram (a detailed description how segments end points of the mass chromatograms are obtained is given under point 6 of the Theory section in the SI). The product of the MCQ values of the corresponding mass chromatograms for each segment of the sample and reference chromatograms were calculated and used as a measure of quality. Parameters were set to select mass chromatograms with MCQ products higher than 0.59 with an upper limit of 30 mass chromatograms per segment.

For a given segment, the product of correlation coefficients was calculated for each pair of selected mass chromatograms using the allowed segment end points for the reference and the sample chromatograms (see point 3 of the Theory section in the SI). The warping procedure uses dynamic programming. In case there is no mass trace selected for a segment, because there is no high-quality mass chromatogram present, no warping is applied for this segment and the cumulated correlation from the previous segment is passed on to the next segment without modification. The algorithm may be improved by requesting a minimum number of selected mass traces for warping and relating the number of selected traces and the product of their MCQ values to the allowed retention time correction, thus tolerating larger retention time shifts for segments containing a high number of high-quality mass traces.

The segmentwise trace selection using MCQ products enables not only the selection of high-quality mass traces containing information about the peaks but also the selection of traces from peaks occurring in both chromatograms. This favors alignments that are based on conserved peaks (compounds) and reduces the risk of misalignments in crowded regions of chromatograms from highly complex samples.

Form of the Benefit Function. The benefit function in COW using one-dimensional information for time alignment is the sum of the Pearson correlation between segments of reference and sample chromatograms. In COW-CODA, this output function is replaced with the correlation product of segments for the selected mass traces. The overall benefit function is therefore the sum of the correlation products. A detailed description of how the correlation product and the corresponding benefit function are obtained is given in Figure S-1 (SI).

Choosing the Reference Chromatogram. Aligning multiple chromatograms requires selection of a reference chromatogram to which the other chromatograms in the data set shall be aligned. To select the best reference, we have calculated the correlation of LC−MS chromatograms based on the reconstructed TIC from all CODA-selected mass traces (CODA-TIC) with 100 different MCQ thresholds between 0.80 and 0.99 (with an interval of 0.0019) across the entire retention time range, during which relevant peaks elute (∼44−130 min for serum data set and ∼13−105 min for the urine data set). For each chromatogram and each studied MCQ value, we calculated the sum of the correlation coefficients of the CODA-TIC profile before time alignment and the chromatogram giving the highest sum of correlation was chosen as the best reference. Consequently, the chromatogram giving the lowest sum was chosen as the worst reference. The selected best and worst reference chromatograms were plotted as a function of MCQ value, and the chromatograms that were most often selected as best and worst references were chosen to perform the time alignment giving a "best" and "worst case" scenario. The complete flowchart of reference chromatogram selection is presented in Figure S-2 (SI).

Global Evaluation of the Time Alignment Quality. We can easily compare the quality of time alignment for single peaks by visual inspection of multiple data sets using internal standards (peptides or metabolites) that were added to the samples. However, it is difficult to judge the overall quality of the time-warping algorithm for complex chromatograms by visualizing only a small number of selected compounds. To overcome this limitation, we have developed a procedure to assess the quality of time alignment based on the calculation of the overlapping peak area between pairs of chromatograms. Peak areas were calculated after applying a modified local baseline subtraction method called $M - N$ rules using $M = 3$ and $N = 8$.[22] An increased quality of time alignment between two chromatograms is reflected in a larger overlapping peak area. The sum of overlapping peak areas from each chromatogram to all of the other chromatograms in the original data set is then compared to the overlapping peak area after applying COW-TIC or COW-CODA. Using this ap-

Table 1. Overview of Preanalytical Parameters and Their Levels in a Fractional Factorial Design Study of Depleted and Trypsin-Digest Serum (Data Set 2)

experiment	run order	blood collection tube	hemolysis	clotting time (h)	freeze−thaw cycles	trypsin digestion	stopping trypsin	stability sample (days)
N1	11	BD368430	low	2	1	1:20	yes	0
N2	3	BD367784	low	2	1	1:100	yes	30
N3	9	BD368430	high	2	1	1:100	no	0
N4	17	BD367784	high	2	1	1:20	no	30
N5	18	BD368430	low	6	1	1:100	no	30
N6	2	BD367784	low	6	1	1:20	no	0
N7	10	BD368430	high	6	1	1:20	yes	30
N8	16	BD367784	high	6	1	1:100	yes	0
N9	8	BD368430	low	2	3	1:20	no	30
N10	15	BD367784	low	2	3	1:100	no	0
N11	13	BD368430	high	2	3	1:100	yes	30
N12	7	BD367784	high	2	3	1:20	yes	0
N13	12	BD368430	low	6	3	1:100	yes	0
N14	6	BD367784	low	6	3	1:20	yes	30
N15	5	BD368430	high	6	3	1:20	no	0
N16	19	BD367784	high	6	3	1:100	no	30
N17	14	BD368430	low	2	1	1:20	yes	0
N18	1	BD368430	low	2	1	1:20	yes	0
N19	4	BD368430	low	2	1	1:20	yes	0

proach, the performance of the time alignment methods can be compared relative to each other. This evaluation method considers a much larger number of chromatographic peaks than the internal standard method. It is, however, not able to judge whether the time alignment method is entirely accurate. In order to check the number of possible misalignments after warping, we visually inspected three peaks per time segment (number of segments were 51, 84, and 41 for data sets 1, 2, and 3, respectively) having the largest average peak intensity in each data set.

MATERIALS AND METHODS

Chemicals. Acetonitrile (ACN) HPLC-S gradient grade (Biosolve; Valkenswaard, The Netherlands), ultrapure water (18.2 MΩ/cm), trifluoroacetic acid (TFA) 99% spectrophotometric grade (Aldrich; Milwaukee, WI), and formic acid (FA) 98−100% pro analysis (Merck; Darmstadt, Germany) were used for reagent and solvent preparation.

Serum Samples. *Cervical Cancer Patients (Data Set 1)*. Serum samples from 10 cervical cancer patients at two different time points (time point A, before treatment with confirmed cancer; time point B, after treatment with no recurrence of the cancer for at least 6 months) were obtained from the Department of Gynecological Oncology (University Medical Centre, Groningen, The Netherlands). The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004. All samples were stored at −80 °C in aliquots. The samples from cervical cancer patients were depleted of the six most abundant serum proteins on a Multiple Affinity Removal column (4.6 × 50 mm, No, 5185-5984, Agilent Technologies) followed by digestion of the remaining proteins with trypsin. Further details about the LC−MS analyses are described in Govorukhina et al. using a quadrupole ion trap mass spectrometer[40]

Factorial Design (Data Set 2). For the factorial design study, serum samples from a healthy female volunteer were obtained from the Department of Gynecological Oncology (University Medical Centre, Groningen, The Netherlands) and stored at −80 °C in aliquots. The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004.

The sample preparation of the serum in this data set was similar to the sample preparation of the serum for data set 1, except for the seven following preanalytical parameters, which were varied at two levels according to a 2_{VI}^{7-3} fractional factorial design (see Table 1). In brief, these factors were type of blood collection tube, hemolysis level, clotting time, number of freeze−thaw cycles, trypsin concentration, deactivation of trypsin after digestion, and stability of the digested sample in the autosampler of the LC−MS system. Sixteen from the 128 possible combinations were selected with 3 repetitions of one condition resulting in a total number of 19 analyses. Digested serum samples were analyzed by LC−MS according to Govorukhina et al. [40]

Urine Samples (Data Set 3). First-void midstream morning urine samples were obtained from 25 pregnant females from a local biobank (Department of Obstetrics and Gynecology of the University Medical Center in Groningen, The Netherlands) and stored frozen at −20 °C. Twenty-five first-void midstream morning urine samples from nonpregnant females were collected in polypropylene containers and kept at 4 °C for a maximum of 1 day, after which they were aliquoted in 10-mL polypropylene tubes and stored at −20 °C. The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004.

Urine samples were thawed, mixed, and acidified with TFA to reach a final concentration of 1%. Samples were left overnight on melting ice and centrifuged to remove precipitate (10 min at 1500*g* and 4 °C). The supernatant was diluted 1:1 with 0.2% FA in 10% ACN and stored at 4 °C until analysis. Supernatants of the acid-precipitated urine samples were analyzed by LC−MS as described by Kemperman et al.[39]

Data Analysis. For processing and multivariate statistical analysis, the original Bruker Daltoniks LC−MS data files were converted to ASCII format with the Bruker Data Analysis software. The ASCII files were transformed into a matrix with the following dimensions: retention time, m/z value and intensity. Data reduc-

tion was performed to combine m/z ratios into 1 amu bins (originally 0.1 amu) by multiplying the original data with a weight-normalized two-dimensional Gaussian weight matrix. This was followed by time alignment using the COW-CODA or the original COW-TIC algorithm. All alignments were done with respect to the best and the worst reference chromatogram. The following parameters were used for the data sets obtained from analyzing trypsin-digested serum (factorial design and cervical cancer, respectively) segment length m: 139 data points (~2.3 min) and 84 data points (1.5 min), dividing each chromatogram into 51 and 84 segments; slack parameter t, 28 and 17 data points. The following parameters were used for the data sets obtained from analyzing acid-precipitated urine: segment length m, 83 data points (~2.2 min); number of segments, 41; slack parameter t, 16 data points.

For each aligned chromatogram, a peak list was generated using $M - N$ rules with M set to 3 and N to 8. The peak lists, generated from all chromatograms (samples), were used to create one common matched peak matrix per study. In order to combine peak lists, one-dimensional peak matching was performed using the sliding window technique, in which the same m/z traces were evaluated for peaks that are proximate in time (step size 0.1 min; search window 1.0 min; maximal accepted standard deviation for all retention times within a group of matched peaks 0.75 min). Missing peak locations were filled with the calculated, background-subtracted local intensity in the respective chromatogram obtained at a location determined from the average m/z and retention time of the corresponding peaks in samples where they were detected. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak (row)−sample (column)−intensity (value) matrix.

All data preprocessing work was done on a personal computer equipped with a +3800-MHz AMD processor and 4 GB of RAM.

RESULTS AND DISCUSSION

Design of the Study. The time alignment algorithm was applied and evaluated with three different data sets obtained from proteomics or metabolomics profiling studies. The first set of analyzed samples was serum-depleted of the six most abundant proteins and trypsin-digested. These samples resulted from a study to discover novel biomarker candidates for cervical cancer, and they are typical for highly complex, shotgun proteomics samples. Previous work had shown that the concentration sensitivity of this method lies in the 0.5 μM range,[40] an area where mainly high-abundant proteins with low biological (interpatient) variability are detected. It is, however, noteworthy that a number of investigators have recently shown that this part of the proteome may also change in a disease-dependent manner due to residual proteolytic activity.[41] This LC−MS data set (referred to as data set 1) was acquired under strict standard operating conditions and thus contained low analytical variability.[40] The factorial design data set (data set 2) was obtained from depleted and trypsin-digested serum from one healthy female volunteer and thus contained no biological variability. However, seven preanalytical factors were deliberately varied, and the effect on the overall proteomics profile was studied at two different levels resulting in considerable analytical variability (see Table 1 for details). Data set 3 was obtained by analyzing acid-precipitated urine, a body fluid containing largely low molecular weight metabolic end products of the organism destined for excretion, from 50 different women. Due to a very high level of biological variability in this data set, the generation of a common aligned peak matrix from different LC−MS analyses proved particularly challenging. As this data set was also acquired under stringent standard conditions, it had low analytical variability.[39] The serum samples contain on average 10 800 features extracted under the specified MN rules, corresponding to 2200−3600 compounds, while the urine samples contain 13 550 extracted features, corresponding to 2700−4500 peaks. The distribution of peaks in the retention time−m/z space is quite uniform for urine samples, while the serum samples show an elliptical distribution from low retention time and low m/z values to high retention time and high m/z values (see Figure S-3 in the SI).

Comparison between the COW-TIC and the COW-CODA Algorithm. The performance of the COW-TIC and the COW-CODA algorithms was evaluated on the basis of annotated peaks originating from added, known compounds as well as by using a global evaluation strategy based on the calculated overlapping peak areas between the reference and the various target chromatograms.

Evaluation Based on Added, Known Compounds. Figure 2 shows extracted ion chromatograms of selected internal standards using the original data (black traces), aligned data after applying the COW-TIC (blue traces), or the COW-CODA algorithm (red traces) using the best reference chromatograms. Visualization of these peaks (spiked peptides for urine samples and horse heart cytochrome *c*-derived tryptic peptide fragments for serum samples) in the original data sets shows that the initial retention time shifts are on the average 0.52 min for data set 1 (cervical cancer), 0.23 min for data set 2 (factorial design), and 0.31 min for data set 3 (urine). The COW-TIC algorithm using a one-dimensional benefit function was only able to align data set 1, which produced well-defined, highly similar TICs across all samples despite the sometimes rather large differences in retention time (up to 2.5 min) between runs (Figure 2a). Results obtained with the COW-CODA algorithm were comparable to COW-TIC in this case.

Although the complexity of the original data set 2 is similar to data set 1, the COW-TIC algorithm was unable to align the chromatograms correctly due to rather dissimilar TIC profiles (large analytical variability due to the deliberate variation of reanalytical parameters). Indeed, in some chromatograms, COW-TIC increased the retention time shift differences compared to the original time differences. The COW-CODA algorithm, on the other hand, resulted in clearly improved alignment resembling the result obtained with data set 1 (Figure 2b). The difference in performance of the algorithms was even more pronounced when trying to align the 50 chromatograms of data set 3 (urine), which contains large biological variability. The COW-TIC algorithm (Figure 2c) lead to a number of obvious misalignments, while COW-CODA resolved the initial misalignments correctly leading to a well-matched set of data, except for the standard peak eluting around 39 min (see Figure S-4 in SI). Even for this peak retention,

(41) Villanueva, J.; Shaffer, D. R.; Philip, J.; Chaparro, C. A.; Erdjument-Bromage, H.; Olshen, A. B.; Fleisher, M.; Lilja, H.; Brogi, E.; Boyd, J.; Sanchez-Carbayo, M.; Holland, E. C.; Cordon-Cardo, C.; Scher, H. I.; Tempst, P. *J. Clin. Invest.* **2006**, *116*, 271–84.

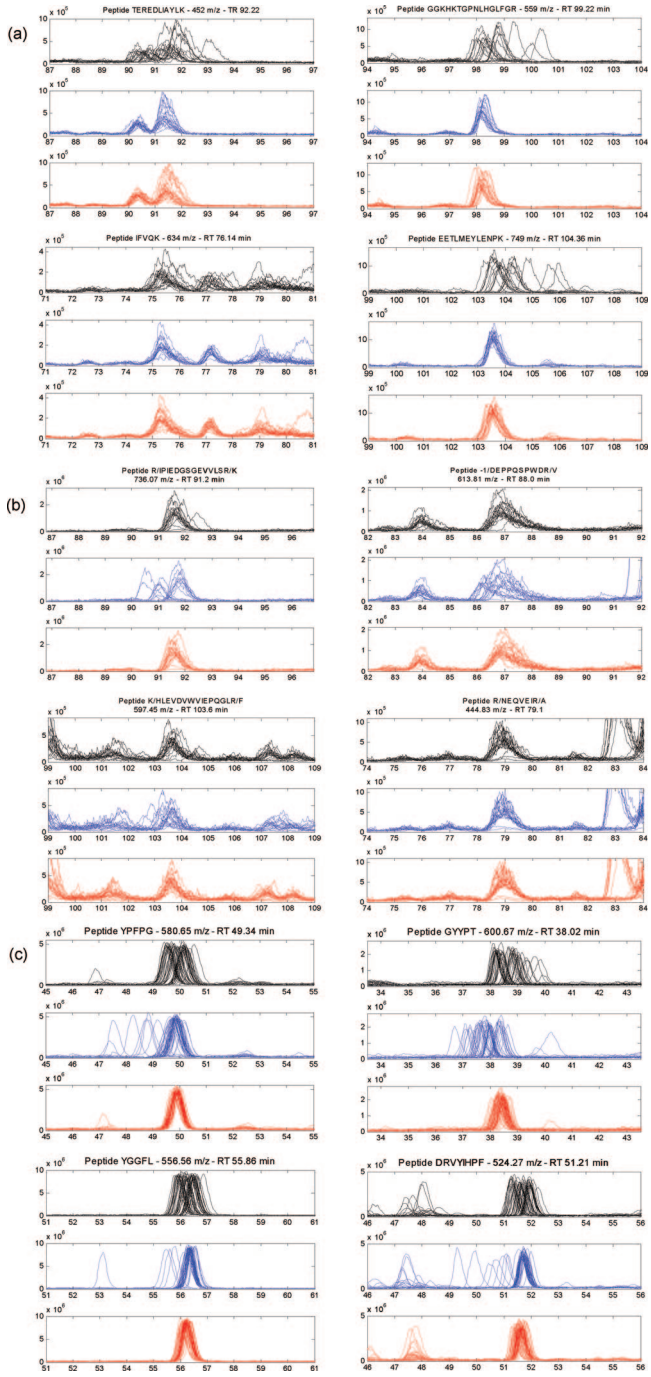

Figure 2. Extracted ion chromatogram of internal standard peptide (a) data set 1 (cervical cancer serum) composed of 20 chromatograms, (b) data set 2 (factorial design serum) composed of 19 chromatograms, and (c) data set 3 (acid-precipitated urine) composed of 50 chromatograms. Each peptide is presented before alignment (top/black), after alignment by COW-TIC (middle/blue), and after alignment by COW-CODA (bottom/red). These time alignment results were obtained using the best references, which were chromatograms 2, 14, and 25 for data sets 1, 2, and 3, respectively. Significant misalignments are observed for data sets 2 and 3 when the COW-TIC algorithm was used, while the COW-CODA algorithm resulted in well-aligned peak clusters.

time shifts were considerably reduced. Thus, only application of the COW-CODA algorithm resulted in clearly improved alignments of all studied data sets. Similar results were obtained when aligning to the worst reference chromatogram (Figure S-4 in the SI) indicating that selection of the reference chromatogram has no effect on the final quality of warping. This simplifies data processing, since one may start the alignment with any chromatogram as the reference.

Global Evaluation of Alignment Quality. The sums of overlapping peak areas between pairs of chromatograms from the same data set were obtained with the original data, after applying the COW-TIC or the COW-CODA algorithm, using the best or the worst reference chromatogram (Figure 3 and Figure S-5 in SI). The COW-CODA algorithm improved time alignment for all three data sets compared to the original, nonaligned data and for data sets 2 and 3 with respect to the COW-TIC algorithm, confirming results obtained with the internal standards. The COW-TIC algorithm led to higher overlapping peak areas than COW-CODA for data set 1, but the difference between the performances of the two algorithms is very small. The slightly better performance of COW-TIC in this case could be due to the fact that CODA-selected single mass traces contain higher levels of noise than the TIC, where the noise of the individual mass traces averages out. The slightly better time alignment using COW-TIC was also observed when inspecting individual peaks of tryptic peptides derived from the added cytochrome *c* (see Figure 2a).

Performance of the COW-CODA algorithm might be improved by applying a higher extent of smoothing on a separate copy of the LC−MS data to be used for calculating the benefit function, while segment-based mass trace selection using CODA is still performed on the original LC−MS data using a small extent of smoothing or no smoothing at all. However, time alignment with COW-CODA was sufficiently accurate for correct peak matching.

The COW-TIC algorithm led to misalignments in the case of the highly variable data set 2 (factorial design; Figure 3b), leading to higher retention time shifts for 16 out of the 19 chromatograms. These results confirm the observations made with peptides derived from added cytochrome *c* (Figure 2b). The COW-CODA algorithm was, on the other hand, able to align the highly variably LC−MS data sets of both the factorial design study and the acid-precipitated urine samples, resulting in higher overlapping peak areas as compared to COW-TIC (Figure 3b and c). Both algorithms improved the overall alignment considerably when compared to the original data for data set 3.

Representing the overlapping peak area of the reference and sample chromatograms per segment reveals the performance differences of the two time alignment algorithms for different segments. Figure S-6 (SI) shows the overlapping peak area of the segments between the best reference and a given sample chromatogram. The figure shows that for most of the segments the two algorithms give similar results. While COW-TIC is performing slightly but not substantially better than COW-CODA for some segments (segments 16−18, 20, 22−23, and 28 show slightly better performance and segment 19 shows moderately better performance), COW-CODA improves alignment of segments 21, 25−27, 29−31, 33, and 38 with substantial improvements for segments 21, 29, and 33.

Retention time correction as a function of retention time obtained with the COW-CODA and COW-TIC algorithm are presented in Figure S-7 (SI) for chromatograms 1 and 14 for data set 2 (factorial design)) showing substantial corrections between

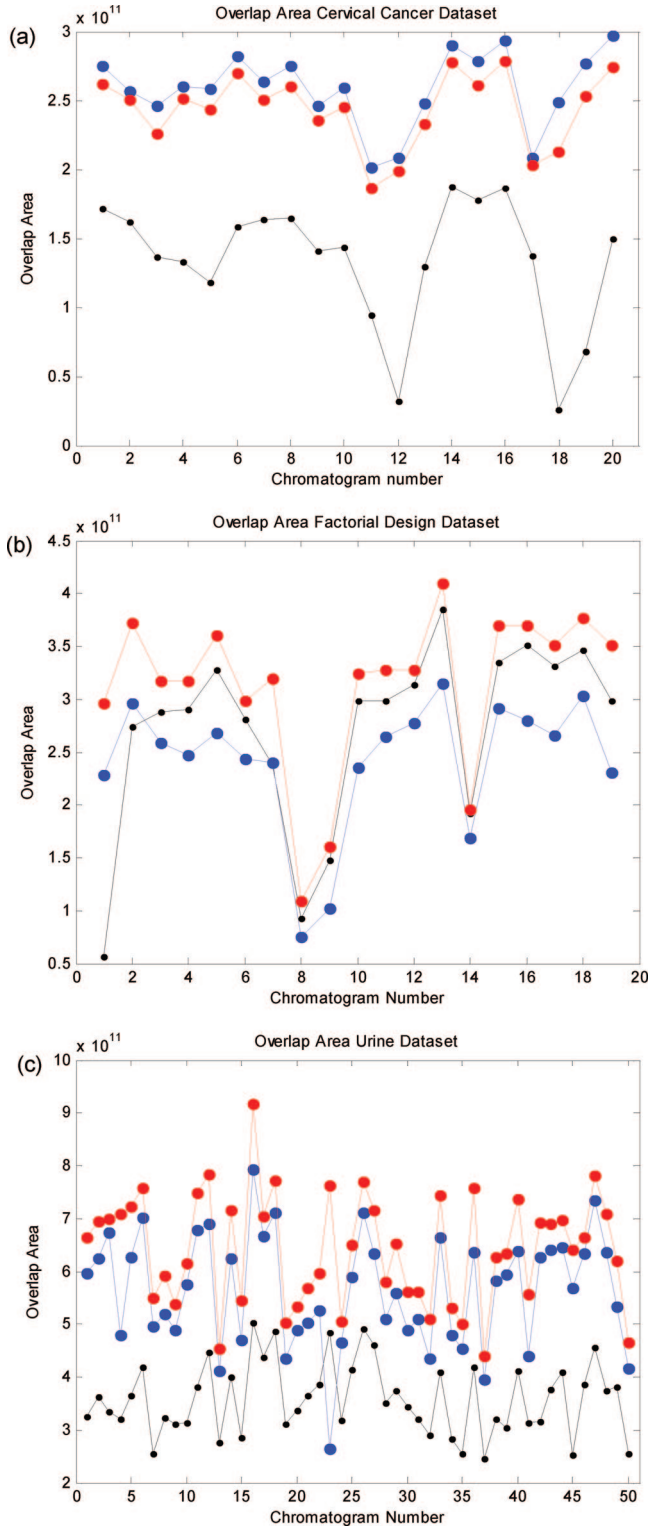

Figure 3. Calculated overlapping peak area after application of *M* − *N* rules to (a) data set 1 (cervical cancer serum), (b) data set 2 (factorial design serum), and (c) data set 3 (acid-precipitated urine). Original nonaligned data (black), aligned with the COW-TIC (blue) or with the COW-CODA algorithm (red). COW-CODA (red) improved the alignment for all three data sets compared to the nonaligned data and with respect to COW-TIC for data sets 2 and 3. These time alignment results were obtained using the best references, which were the chromatograms 2, 14, and 25 for data sets 1, 2, and 3, respectively.

retention times of 110 and 125 min and the superior performance of the COW-CODA algorithm in this difficult region (see Figure S-7, SI).

Effect of Reference Chromatograms. It may be argued that all COW-based algorithms depend strongly on the selected reference chromatogram (see Figure S-8 of the SIand the Materials and Methods section on how reference chromatograms were selected). In order to study this factor in more detail, we compared the overlapping peak areas (Figure S-5, SI) confirming earlier results that showed that the overall quality of alignment using the COW-CODA algorithm depends little on the reference chromatogram. However, small differences were observed in data set 1 for chromatograms 18 and 19 and for a few chromatograms in data set 3 (Figure S5-a and c, SI). The stability of the algorithm for data sets 1 and 2 with respect to the choice of the reference chromatogram increases our confidence that the extent of misalignment with any of the chosen reference chromatograms is small. We assume that a large number of misalignments would have resulted in variable overlapping peak areas depending on the selected reference chromatogram. The fact that the COW-CODA algorithm functions correctly and independently of the chosen reference chromatogram greatly simplifies automation of the time alignment step as part of the overall data processing pipeline for chromatograms obtained from depleted, trypsin-digested serum samples. The slight dependency of the final alignment on the chosen reference chromatogram for data set 3 shows, however, the need for algorithms that select the best reference for data sets with high biological variability. Calculating the overlapping peak area appears to be a suitable way to do so.

Assessing the Inherent Variability of the Data Sets and the Required Processing Time. With a properly aligned common peak matrix, it is possible to assess the variability of the different data sets. A first evaluation of the variability of data sets 1−3 was performed by plotting the number of detected peaks against the relative standard deviation (RSD) (Figure 4). Based on considerations discussed before, it is expected that the RSD histogram obtained from the aligned peak matrix of the depleted and trypsin-digested serum samples from cervical cancer patients (data set 1) would be narrower relative to histograms obtained from the factorial design or urine data sets. Figure 4 shows further that the biological variability of urine not only increases the maximum RSD but also results in a broader RSD distribution compared to serum data sets, indicating that peaks have rather high variability in intensity in urine. This stands in contrast to the serum-derived data set 2, where variation is only due to preanalytical parameters, which increases only the mode of RSD without broadening the RSD distribution. Accurate time alignment thus allows us to measure the variability that is inherent to the data sets resulting from analytical and biological variations. Such histograms may be used to perform statistical simulations, e.g., for power calculations, or to test the performance of different variable selection and classification algorithms.

Alignment of one pair of chromatograms obtained with serum samples (~7100 scans) took ~22 min using COW-CODA and 0.8 min using COW-TIC with an ordinary PC as described in Materials and Methods. The processing time for chromato-

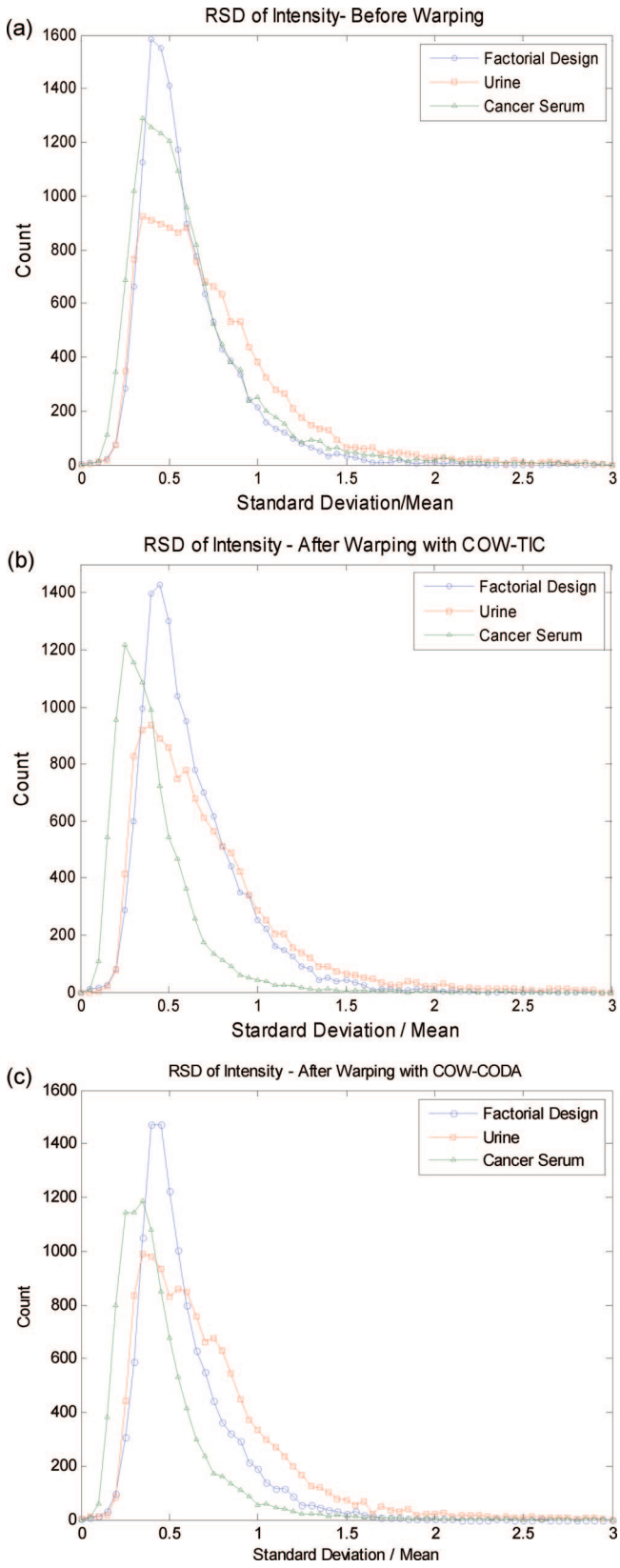

Figure 4. Standard deviation of the intensity of peaks selected with $M - N$ rules (M = 3; $N = 8$) in data sets 1 (cervical cancer serum) (green), 2 (factorial design serum) (blue), and 3 (acid-precipitated urine) (red) divided by the mean before warping (top) and after warping with the COW-CODA algorithm (middle) and after warping with COW-TIC (bottom). These time alignment results were obtained using the best references, which were chromatograms 2, 14, and 25 for data sets 1, 2, and 3, respectively.

grams obtained from urine samples (~3450 scans) was 5.35 min for COW-CODA and 0.2 min for COW-TIC. The ~27 times higher processing time of COW-CODA is mainly due to the higher I/O, since the algorithm operates on the full LC−MS data set contrary to the COW-TIC algorithm, which is only using a single trace, the TIC, for the time alignment procedure. Translating the COW-CODA algorithm into a low-level programming language will increase the processing speed, which may be further reduced by parallel processing of pairs of multiple chromatograms.

CONCLUSIONS

We describe an improved time alignment algorithm based on COW combined with a segmentwise selection of high-quality mass chromatograms using CODA. This algorithm is effective in aligning highly complex proteomics and metabolomics LC−MS data sets containing different levels of analytical and biological variability. The novel algorithm outperforms the original COW-TIC algorithm in the case of data sets containing either a high level of analytical (factorial design study) or biological (acid-precipitated urine) variability.

The presented algorithm uses a two-dimensional benefit function in order to discriminate between peaks with different m/z values eluting at similar or identical retention times and to select mass traces sharing common, high-quality peaks. We have observed only very minor misalignments after visually inspecting a few hundred extracted ion chromatograms for each of the analyzed samples. This strongly supports the conclusion that the COW-CODA algorithm results in a very high quality of time alignment and that the studied data sets contain a sufficient number of common peaks in each time segment to drive the alignment procedure to a global optimum.

Another advantage of COW-CODA is that only a few parameters need to be selected and optimized depending on the data set (e.g., adapting the chosen segment length to the observed chromatographic peak width) and that the final quality of alignment is rather independent of the initially chosen reference chromatogram.

This article presents results from low-resolution MS data. However, COW-CODA is also applicable to high-resolution data using Gaussian smoothing and data reduction to 1 amu in the mass dimension, since the algorithm only calculates the corrected retention time of the mass scans (Figure S-9 in SI).This "binning" does not affect the ultimate resolution of the MS data, since the warped retention time values of the mass scans can be applied to data with the original resolution.

We are presently working on combining CODA with other algorithms to investigate whether this is a general approach to improve time alignment of highly complex LC−MS data sets. Other mass chromatogram selection algorithms may be used as alternatives to CODA,[42] although CODA performed well in the tested cases.

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. The work described in this publication was supported by the project BioRange 2.2.3 from The Netherlands Proteomics and The Netherlands Bioinformatics Center.

(42) Nyangoma, S. O.; van Kampen, A. A.; H, C.; Reijmers, T. H.; Govorukhina, N. I.; van der Zee, A. G. J.; Billingham, L. J.; Bischoff, R.; Jansen, R. C. *Stat. Appl. Genet. Mol. Biol.* **2007**, *6*, 23.

SUPPORTING INFORMATION AVAILABLE

Detailed theory of COW-CODA with glossary of mathematical terms. Eleven figures. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review May 4, 2008. Accepted June 29, 2008.

AC800920H