Journal of Statistical Software

January 2007, Volume 18, Issue 4. http://www.jstatsoft.org/

amsrpm: Robust Point Matching for Retention Time Alignment of LC/MS Data with R

Marc Kirchner
University of Heidelberg

Benjamin Saussen
University of Heidelberg

Hanno Steen
Children's Hospital Boston

Judith A. J. Steen
Children's Hospital Boston

Fred A. Hamprecht
University of Heidelberg

Abstract

Proteomics is the study of the abundance, function and dynamics of all proteins present in a living organism, and mass spectrometry (MS) has become its most important tool due to its unmatched sensitivity, resolution and potential for high-throughput experimentation. A frequently used variant of mass spectrometry is coupled with liquid chromatography (LC) and is denoted as "LC/MS". It produces two-dimensional raw data, where significant distortions along one of the dimensions can occur between different runs on the same instrument, and between instruments. A compensation of these distortions is required to allow for comparisons between and inference based on different experiments. This article introduces the amsrpm software package. It implements a variant of the Robust Point Matching (RPM) algorithm that is tailored for the alignment of LC and LC/MS experiments. Problem-specific enhancements include a specialized dissimilarity measure, and means to enforce smoothness and monotonicity of the estimated transformation function. The algorithm does not rely on pre-specified landmarks, it is insensitive towards outliers and capable of modeling nonlinear distortions. Its usefulness is demonstrated using both simulated and experimental data. The software is available as an open source package for the statistical programming language R.

Keywords: registration, alignment, sequence alignment, dynamic time warping, monotone regression, retention time, elution, LC/MS, robust point matching, chromatogram warping.

1. Introduction

In many Proteomics mass spectrometry (MS) experiments, mass analyzers are operated in

line with liquid chromatography (LC) systems. LC systems use an organic solvent gradient to separate a biological sample based on the chemical properties of its constituents before subjecting it to ionization and mass analysis, thus decreasing sample complexity by improving analyte separation. Research on many biological questions involves comparative analysis of LC/MS data acquired from samples which have been exposed to different environmental conditions. An undesirable feature of LC separation techniques is the introduction of a random time shift, i.e. a nonlinear distortion in the retention time domain, leading to significant run-to-run variability, thus compromising comparability of experiments. Hence, in order to allow for differential analysis, the acquired data has to be aligned/registered.

Alignment/registration problems have a long history in chromatography: straightforward shift correction based on regression and/or correlation methods include Reiner *et al.* (1979) who proposed piecewise linear interpolation. Wang and Isenhour (1987) introduced *Dynamic Time Warping (DTW)* into the chromatography field. Originally developed by Itakura (1975) and Sakoe and Chiba (1978) for speech signals, DTW calculates a nonlinear mapping between two samples using constrained dynamic programming (Dreyfus 2002). Recent applications and analyses have been carried out by Pravdova *et al.* (2002) and Ramaker *et al.* (2003), all of which also focus on *Correlation Optimized Warping (COW)*, a signal segment-wise correlation optimization procedure introduced by Nielsen *et al.* (1998). Eilers (2004) introduced *Parametric Time Warping (PTW)*, a landmark-free iterative least-squares approach estimating three parameters for a quadratic warping function. Although susceptible to noise effects, the hierarchical clustering method of Tibshirani *et al.* (2004) was the first approach to overcome the missing correspondence problem, subsequently addressed in the fuzzy warping approach of Walczak and Wu (2005). Listgarten *et al.* (2005) propose the *Continuous Profile Model (CPM)*, a generative model where each observed time series is considered a sub-sampled realization of a latent trace. Sample registration in the full LC/MS domain has been carried out by Listgarten *et al.* (2007), following their earlier CPM approach but cutting down the computational burden with a spline approximation technique. Fischer *et al.* (2006) propose a semi-supervised alignment algorithm incorporating chemical information available from tandem mass spectrometry (MS^2) into a nonlinear ridge regression model. They subsequently use a least predictive variance self-training scheme to iteratively improve their warping estimates. Krebs *et al.* (2006) advocate peak picking prior to chromatogram registration by landmark selection and assessment in high intensity elution regions. Comprehensive reviews on the topic include Tomasi *et al.* (2004) and van Nederkassel *et al.* (2006).

The *Robust Point Matching (RPM)* procedure by Chui and Rangarajan (2002) was developed in the context of medical image processing where image registration problems are commonly encountered. The difficulty of this task is largely determined by the amount of information that is available: if landmarks are known, finding the *warping function* often is straightforward, if correspondences (homologies) still have to be determined, i.e. if one has to deal with unlabeled features, the problem becomes considerably harder. Also, when considering chromatograms as point sets (e.g. after peak picking), their cardinality need not be equal, hence calling for outlier detection and rejection methods. Based on earlier work of Rangarajan *et al.* (1996), and Rangarajan *et al.* (1999), RPM provides functionality to deal with each of these problems. Inheriting from the *Iterative Closest Point (ICP)* algorithm, the *Softassign Procrustes Matching* (Rangarajan *et al.* 1997), and employing *deterministic annealing*, RPM is an iterative EM-style correspondence and transformation estimation scheme using gradually stabilizing quasi-probabilistic correspondences which serve as basis for the calculation of

gradually refining *Thin Plate Spline* (*TPS* Bookstein 1989) warping functions.

The **amsrpm** package has been developed with quantitative mass spectrometry data processing in mind. In this field, to date, available registration/alignment information is often limited to retention time, mass-to-charge ratio and peak intensity. In general, there is no MS^2 information available, due to the necessity of high sampling rates, which are severely downgraded by MS^2 sampling. It is important to note that this is the very reason that forces us to use notions of "proximity" to estimate correspondences instead of using chemical information that can help establish true correspondences.

Section 2 introduces the data format generally associated with LC/MS experiments. Section 3 details the **amsrpm** algorithm and its relations to RPM. Section 4 shows the application of the algorithm to a simulated and a real-world data set. Finally, conclusions and an outlook are offered in section 5.

2. Data

LC/MS experiments yield two-dimensional intensity distributions. The mass analyzer records count/intensity values which are indexed by their elution time (the time required by an analyte to pass through the LC column) and their molecular mass/electric charge ratio. Due to the sheer amount of data ($\geq 2 \cdot 10^6$ points in our experimental data set Pho4), and since a single analyte typically accounts for more than one detected event, raw peak data are usually preprocessed by a *peak picking* method, aimed at extracting the relevant peaks from the raw data, thus producing a succinct, more manageable LC/MS *peak list*. A mere LC recording can be obtained by integrating the raw data over the mass/charge domain, yielding the *Total Ion Current (TIC)* chromatogram. In both cases, we can consider a data set as a point set

$$A = \{\mathbf{a}_i | i = 1, \ldots, N_A\}, N_A \in \mathbb{N}. \tag{1}$$

For TIC chromatograms each point is defined by its retention time a_{i1} and TIC intensity a_{i2} (cf. Figures 1 and 2). With LC/MS peak lists, each point has coefficients a_{i1}, a_{i2}, and a_{i3} being retention time, mass/charge ratio, and intensity, respectively (cf. Figure 3). Hence, registration/alignment deals with the point sets A and B, where A is defined as in eq. (1), $B = \{\mathbf{b}_i | i = 1, \ldots, N_B\}, N_B \in \mathbb{N}$, $\mathbf{a} = (a_{1i}, a_{2i}, \ldots, a_{di})^T$, and $\mathbf{b} = (b_{1i}, b_{2i}, \ldots, b_{di})^T$, with $d \in \{2, 3\}$.

The **amsrpm** package can be used for the registration of either chromatograms or peak lists acquired in different experiments. To demonstrate both functionalities, section 4 shows results for both types of data. While the TIC chromatograms used in the demonstration were extracted from the raw LC/MS data, we would like to point out that the use of full peak lists is recommended whenever possible, since these contain much more information than the TIC chromatograms obtained from them.

3. Methods

In analogy to Chui and Rangarajan (2002), we can view the alignment problem as a point set registration task split up into repetitive EM-style correspondence and transformation estimation steps, coupled to a deterministic annealing scheme. On convergence, this yields a

nonlinear transformation, providing an estimate for transforming B onto A: $\hat{A} = \mathcal{F}(B)$, with $|\hat{A}| = |B| = N_B$.

In the alignment of either TIC chromatograms or LC/MS peak lists, the observed intensities can help in the estimation of correspondences, but should not be modified in the alignment, since it is the information that is of interest for a differential analysis. In the case of LC/MS experiments, shifts in the mass/charge domain between different experiments are non-existent or negligible if the mass analyzer has been properly calibrated. Hence, it is sufficient in either case to limit oneself to the estimation of a nonlinear distortion $f(\cdot)$ in the retention time domain, yielding a one-dimensional regression problem.

Similar to Fischer *et al.* (2006) and Eilers (2004), we consider chromatographic retention time distortions to generally be nonlinear but smooth and thus confine the transformation function space to smooth functions. More precisely, defining D^n as the nth temporal derivative operator, we expect $\int D^2 f(x)dx$ to be bounded. Also, distortions reversing the order in which analytes elute off the column are rare, motivating a monotonicity constraint on the warping function $f(\cdot)$. In the following, we will deal with each of these points in turn.

Correspondence Estimation. We adapt *fuzzy correspondences* from Chui and Rangarajan (2002) similar to Walczak and Wu (2005) leading to a $N_A \times N_B$ correspondence matrix $\mathbf{M} = \{m_{ij}\}$ with

$$m_{ij} = \frac{1}{T} \exp\left\{ -\frac{1}{2T} \psi(\mathbf{a}_i, \mathcal{F}(\mathbf{b}_j)) \right\}, \tag{2}$$

where m_{ij} is the correspondence coefficient between $\mathbf{a}_i$ and $\mathbf{b}_j$. The dissimilarity function $\psi(\cdot,\cdot)$ is defined differently for the TIC chromatogram and full LC/MS case: for TIC registration we include derivative information into the estimation process. Therefore

$$\psi(\mathbf{a}_i, \hat{\mathbf{a}}_j) = (\mathbf{a}_i - \hat{\mathbf{a}}_j)^T \mathbf{W}_2 (\mathbf{a}_i - \hat{\mathbf{a}}_j) + \kappa_{slope} |[D\mathbf{a}]_i - [D\hat{\mathbf{a}}]_j| + \kappa_{sign} \mathrm{H}([D\mathbf{a}]_i [D\hat{\mathbf{a}}]_j) \tag{3}$$

with $\hat{\mathbf{a}}_j = \mathcal{F}(\mathbf{b}_j)$, and $\mathrm{H}(\cdot)$ the Heavyside step function. The weight matrix $\mathbf{W}_2 = \mathrm{diag}(w_1, w_2)$ allows for relative weighting of retention time and intensity distances, κ_{slope} and κ_{sign} control the contribution of derivative information to the dissimilarity measure. With $\kappa_{slope} > 0$ and $\kappa_{sign} > 0$, eq. (3) will favor correspondences between points if their gradients are similar and of equal sign. This proved to be an important addition, significantly improving alignment outcomes.

Registration of LC/MS peak lists deals with arbitrarily spaced point sets and will most frequently be carried out on preprocessed, peak-picked data, where no sensible derivative information is available. We define a weighted squared dissimilarity

$$\psi(\mathbf{a}_i, \hat{\mathbf{a}}_j) = (\mathbf{a}_i - \hat{\mathbf{a}}_j)^T \mathbf{W}_3 (\mathbf{a}_i - \hat{\mathbf{a}}_j) \tag{4}$$

with $\hat{\mathbf{a}}_j = \mathcal{F}(\mathbf{b}_j)$ and $\mathbf{W}_3 = \mathrm{diag}(w_1, w_2, w_3)$ weighting retention time, mass/charge ratio and intensity contributions. Thus, while mass/charge ratio and intensity contribute to the correspondence estimation, the warping function accounts for variation along the retention time axis only.

In order to deal with outliers (points for which no plausible correspondence can be established), we augment $\mathbf{M}$ with an outlier column and row, yielding an $(N_A + 1) \times (N_B + 1)$ matrix $\mathbf{M}_0$. In contrast to Walczak and Wu (2005), we set outlier correspondences to a constant value, independent of the number of points. After Sinkhorn standardization (Sinkhorn 1996; Walczak

and Wu 2005) $\mathbf{M}_0$ holds quasi-probabilistic (in the sense that all entries were subjected to iterative row and column normalization and now sum to one) correspondence estimates, subsequently used for the estimation of the warping function.

Estimation of the optimal smooth monotone warping function. Given the correspondence matrix $\mathbf{M} = \{m_{ij}\}$, estimation of the optimal warping function $f(\cdot)$ can be carried out using a smooth monotone regression approach, as detailed in Ramsay and Silverman (2005). We expand an unconstrained function $w(\cdot) \in \mathbb{R}$ in terms of a set of B-spline basis functions $\phi_k(\cdot)$, yielding

$$w(x) = \sum_k c_k \phi_k(x), \tag{5}$$

and define a roughness-penalized fitting criterion for monotone smoothing as

$$\min_w \left\{ (\hat{\tilde{\mathbf{t}}} - \beta_0 - \beta_1 \{D^{-1} e^{D^{-1} w}\}(\mathbf{t}))^T \mathbf{R} (\hat{\tilde{\mathbf{t}}} - \beta_0 - \beta_1 \{D^{-1} e^{D^{-1} w}\}(\mathbf{t})) + \lambda \int w^2(\tau) d\tau \right\}. \tag{6}$$

The vector $\hat{\tilde{\mathbf{t}}}$ holds the expected value of the retention times of the warped point set $\hat{A} = \mathcal{F}(B)$, i.e. $\hat{\tilde{\mathbf{t}}} = (\hat{\tilde{a}}_{11}, \hat{\tilde{a}}_{21}, \ldots, \hat{\tilde{a}}_{N_B 1})^T$, with $\hat{\tilde{a}}_{j1} = \sum_{i=1}^{N_A} m_{jk} a_{k1}$. The retention times of the original point set B are given in $\mathbf{t} = (b_{11}, b_{21}, \ldots, b_{N_B 1})^T$. $\mathbf{R}$ is a symmetric positive definite matrix allowing for unequal weighting of squares and products of residuals. With known residual covariance matrix $\mathbf{\Sigma}$, we have $\mathbf{R} = \mathbf{\Sigma}^{-1}$, otherwise the assumption of uncorrelated errors and treating $\mathbf{R}$ as a diagonal weight matrix is sensible. The parameters $\beta_k, k = 0, 1$ need to be estimated from the data, see Ramsay (1998) for details. By default, **amsrpm** places spline knots at each observation for LC/MS peak list data and on equidistantly sampled retention time points for TIC chromatogram data. This behavior can be controlled by the user.

Minimizing eq. (6) with respect to the spline coefficients c_k results in a numerical optimization problem that can be solved using the Newton-Raphson algorithm. In **amsrpm** the transformation estimation is carried out using the `smooth.monotone(·)` function from the **fda** package by Ramsay and Wickham (2006).

Annealing procedure. Chui and Rangarajan (2002) introduce a deterministic annealing scheme, aimed at gradually allowing more flexible transformations by adapting λ in each iteration using $\lambda_l = \lambda_{init} T$, with T as annealing temperature. The **amsrpm** implementation follows this idea.

4. Results

This section describes the application of the **amsrpm** package to two TIC chromatogram alignment problems and an LC/MS peak list alignment task. We begin each of the subsections with a data description followed by a short comment on how the alignment was obtained.

Simulated TIC distortion data. During **amsrpm** development we created a simulation dataset by artificially distorting real-world TIC chromatograms. The distortion was simulated by uniform drawing of elution time values from the retention time domain of the original signal. We added a time shift (assuming a normal distribution centered around the previously drawn value), and applied a smooth monotone regression against the original retention time values. Final simulation chromatograms were obtained by adding Poisson noise on TIC intensity as well as zero-mean normal noise on retention time. The simulation data were used for evaluation purposes and we show a sample alignment result in Figure 1.

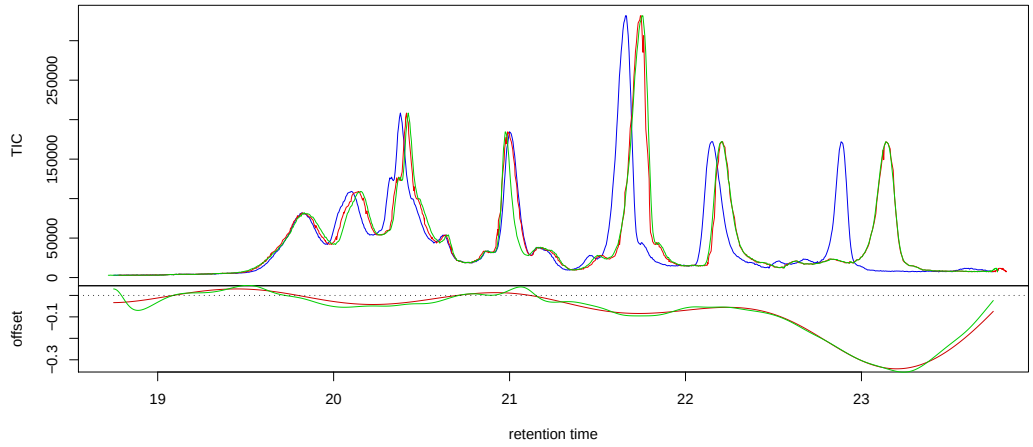

Figure 1: Top panel: TICs of the undistorted sample (blue), after distortion (red), and TIC of the sample after application of the warping estimate with **amsrpm** (green). Bottom panel: Time shifts for the simulated distortion (red) and the **amsrpm** warping estimate (green).

Alignment. Acquiring the alignment estimate is a straightforward procedure: the function `ams.rpm.register.tic(·)` takes two matrices `x` and `y`, with one point per row and columns corresponding to retention time and intensity and outputs the calculated fit. Parameters allow further customization.

```
R> result <- ams.rpm.register.tic(
                x = pho4_tic[[1]],
                y = simchrom,
                kappa = c(0.01, 0.1),
                max.points = 150,
                subsample.method = "equidistant")
```

In this case we have chosen to reweight the derivative sign contribution, to limit the number of points used in the transformation estimation to 150, and to use equidistant subsampling of the TIC. The results are summarized in two panels in Figure 1: plots in the upper panel show TICs of the undistorted sample (blue) and the warped sample (red), and the warping estimation result applied to the undistorted TIC (green). The bottom panel visualizes the time shifts for the simulated distortion (red) and our estimate (green).

Parameter selection. With simulation data available we were able to analyze TIC alignment quality as a function of the weighting parameters κ_{slope} and κ_{sign}. The alignment algorithm was applied to $n = 100$ simulation data sets, with $p = 25$ different parametrizations, where $\kappa_{slope} \in \{0.01, 0.032, 0.1, 0.32, 1\}$ and $\kappa_{sign} \in \{0.001, 0.0032, 0.01, 0.032, 0.1\}$ (logarithmic scales). A summary of results is shown in Table 1. Minimal training error was attained for $\kappa_{slope} = 0.01$ and $\kappa_{sign} = 0.1$.

Pho4 dataset. The Pho4 dataset was initially acquired with relative quantitation of protein phosphorylation in mind. Steen *et al.* (2005) monitored Pho80/85-induced changes in phosphorylation stoichiometry of the yeast transcription factor Pho4. Pho4 was isolated at

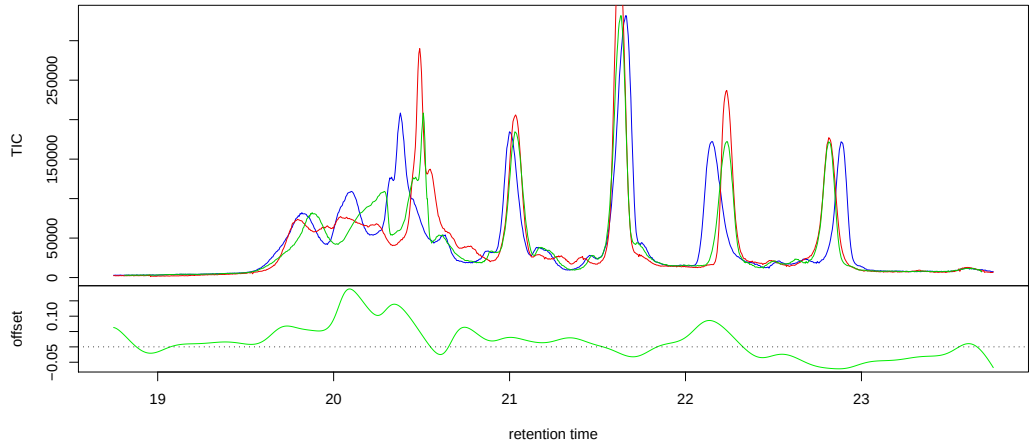

Figure 2: Top panel: **amsrpm** alignment of the Pho4 TIC data, repeats 2 (blue) and 3 (red). The warping result is shown in green. Bottom panel: estimated time shifts.

different time points during a kinase assay. After in-gel tryptic digestion, the samples were analyzed in replicate by LC/MS using a QSTAR XL mass spectrometer (Applied Biosystems/MDS Sciex, Concord, Canada) at MS-acquisition times of 150 ms. The raw data was converted to text files using in-house software. TIC chromatograms were calculated by integration over the mass/charge domain and are provided in the data directory of the **amsrpm** package. We used the *MarkerView* Software (Applied Biosystems/MDS Sciex) for peak picking the raw LC/MS data and to obtain LC/MS peak lists.

Pho4 TIC chromatogram alignment. Alignment of Pho4 TIC chromatogram data proceeds analogously to alignment of the simulated data,

```
R> data("pho4_tic")
R> res <- ams.rpm.register.tic(
            x = pho4_tic[[1]],
            y = pho4_tic[[2]],
            kappa = c(0.01, 0.1),
            max.points = 150,
            subsample.method = "equidistant")
R> ams.plot(res, pho4_tic[[1]], pho4_tic[[2]])
```

yielding the alignment shown in Figure 2. Again, the two original TIC chromatograms are given in blue and red, the warping result in green. It is obvious that this is a more difficult example, with good alignment but problematic regions (especially in the $19.8-20.4$min range, where TIC structure is insufficient to provide a reliable result).

Pho4 LC/MS peak list registration. For LC/MS peak list alignment **amsrpm** provides the `ams.rpm.register.lcms(·)` function:

```
R> data("pho4_lcms")
R> res <- ams.rpm.register.lcms(
```

```
            x = pho4_lcms[[1]],
            y = pho4_lcms[[2]],
            max.points = 200,
            tie.method = "noise")
R> ams.plot(res, pho4_lcms[[1]], pho4_lcms[[2]])
```

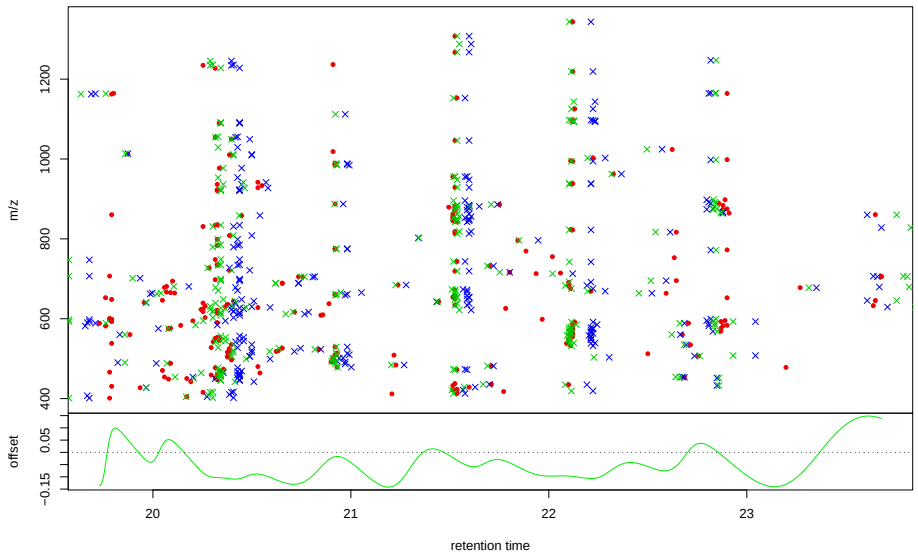

Figure 3: LC/MS alignment. Top panel: Pho4 LC/MS data at $t = 5$min. Result (green) of aligning repeat 2 (blue) with repeat 3 (red). Bottom panel: estimated warping shifts.

Here we chose to limit the number of points contributing to the smooth monotone regression to 200, and to numerically stabilize the fit by adding a tiny amount of noise to the spline knot positions. The top panel of Figure 3 shows the positions of picked peaks constituting the point set B (crosses) before (blue) and after (green) alignment with the point set A (red circles). Difficult alignment regions (neglecting border effects) include two peak clusters around 22.9min, exhibiting slight m/z-dependent shift properties. The same phenomenon can be observed at $\sim$ 20.4min, augmented by imperfect correspondence estimation of the peaks within the cluster at $\sim 600 m/z$. Despite these problems, and recalling that Figures 2 and 3 are based on highly similar data, the benefit of working with LC/MS peak lists as compared to TIC data is obvious.

5. Conclusions and outlook

Conclusions. We present an R package for unsupervised nonlinear alignment/registration of TIC chromatograms and LC/MS peak list data sets. Its capability of automatically finding tentative landmarks makes it well suited for samples of low- to medium complexity where no additional chemical information is available through MS^2 sampling. Imposing physically motivated constraints such as monotonicity leads to sensible and useful warping estimates, and significantly improves run-to-run comparability.

Open Problems and Outlook. The main limitation of **amsrpm** currently is the computational burden involved in repetitively calculating the distance matrix and the smooth monotone regression estimate. The latter is based on **fda** (Ramsay and Wickham 2006), with a native R Newton-Raphson implementation. The package authors themselves note that this is suboptimal in terms of speed. We are currently working at eliminating this shortcoming.

Alignment behavior is controlled by the parameters w_1, w_2, κ_{slope}, and κ_{sign} in the TIC case, and by w_i, $i = 1, 2, 3$ in the LC/MS peak list case. Optimal settings for these parameters depend on the problem at hand and **amsrpm** does currently not provide means to obtain these estimates.

A general limitation of global retention time alignment approaches is their inability to cope with situations in which analytes exhibit mass/charge ratio-dependent run-to-run retention time variability. This also holds for **amsrpm**.

	κ_{slope}	κ_{sign}	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$
1	0.01	0.00	0.72	2.10	22.78
2	0.01	0.00	0.75	2.01	28.04
3	0.01	0.01	0.68	1.55	16.62
4	0.01	0.03	0.51	1.49	18.11
5	0.01	0.10	0.46	**1.24**	13.68
6	0.03	0.00	1.37	3.00	25.01
7	0.03	0.00	0.92	2.63	35.31
8	0.03	0.01	0.95	2.32	17.97
9	0.03	0.03	0.84	2.19	20.14
10	0.03	0.10	0.95	2.34	17.10
11	0.10	0.00	1.53	3.92	24.78
12	0.10	0.00	1.75	3.84	28.09
13	0.10	0.01	1.60	3.85	24.61
14	0.10	0.03	1.37	3.22	17.81
15	0.10	0.10	1.18	3.17	18.93
16	0.32	0.00	2.46	5.86	23.01
17	0.32	0.00	2.74	6.33	35.35
18	0.32	0.01	2.53	7.02	42.76
19	0.32	0.03	2.31	6.29	45.04
20	0.32	0.10	2.28	5.29	23.04
21	1.00	0.00	4.43	10.21	38.09
22	1.00	0.00	3.46	8.77	50.46
23	1.00	0.01	3.81	9.58	40.93
24	1.00	0.03	3.59	7.50	22.45
25	1.00	0.10	3.46	7.04	32.06

Table 1: Quartiles of squared lack of fit of estimated time shifts depending on $\kappa_{slope} \in \{0.01, 0.032, 0.1, 0.32, 1\}$ and $\kappa_{sign} \in \{0.001, 0.0032, 0.01, 0.032, 0.1\}$ parametrization in TIC registration. Quartiles have been calculated over 100 simulation sets for each parametrization.

Acknowledgments

We are indebted to Linus Görlitz, Florian Lenz, Michael Kelm, and Björn Menze for fruitful discussions. The authors gratefully acknowledge support by the Hans L. Merkle foundation (M.K.), the Baden-Württemberg-Scholarship (B.S.), the Children's Hospital Trust (H.S. and J.A.J.S.), and the Robert Bosch GmbH (F.A.H.). We would also like to thank two anonymous reviewers for their helpful comments and critique.

References

Bookstein FL (1989). "Principal Warps: Thin Plate Splines and the Decomposition of Deformations." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 567–585.

Chui H, Rangarajan A (2002). "A New Point Matching Algorithm for Non-rigid Registration." *Technical report*, Department of Computer and Information Science & Engineering, University of Florida.

Dreyfus S (2002). "Richard Bellman on the Birth of Dynamic Programming." *Operations Research*, **50**(1), 48–51.

Eilers PHC (2004). "Parametric Time Warping." *Analytical Chemistry*, **76**, 404–411.

Fischer B, Grossmann J, Roth V, Gruissem W, Baginsky S, Buhmann JM (2006). "Semi-supervised LC/MS Alignment for Differential Proteomics." *Bioinformatics*, **22**(14), 132–140.

Itakura F (1975). "Minimum Prediction Residual Principle Applied to Speech Recognition." *IEEE Transactions on Acoustics, Signal, and Speech Processing*, **23**, 77–92.

Krebs MD, Tingley RD, Zeskind JE, Holmboe ME, Kang JM, Davis CE (2006). "Alignment of Gas Chromatography-Mass Spectrometry Data by Landmark Selection from Complex Chemical Mixtures." *Chemometrics and Intelligent Laboratory Systems*, **81**, 74–81.

Listgarten J, Neal RM, Roweis ST, Emily A (2005). "Multiple Alignment of Continuous Time Series." *Neural Information Processing Systems*, **17**, 817–824.

Listgarten J, Neal RM, Roweis ST, Wong P, Emili A (2007). "Difference Detection in LC-MS Data for Protein Biomarker Discovery." *Bioinformatics*, **23**, 198–204.

Nielsen NPV, Carstensen JM, Smedsgaard J (1998). "Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping." *Journal of Chromatography A*, **805**, 17–35.

Pravdova V, Walczak B, Massart DL (2002). "A Comparison of Two Algorithms for Warping of Analytical Signals." *Analytica Chimica Acta*, **456**(1), 77–92.

Ramaker HJ, van Sprang ENM, Westerhuis JA, Smilde AK (2003). "Dynamic Time Warping of Spectroscopic BATCH Data." *Analytica Chimica Acta*, **1–2**, 133–153.

Ramsay J (1998). "Estimating Smooth Monotone Functions." *Journal of the Royal Statistical Society B*, **60**, 365–375.

Ramsay JO, Silverman BW (2005). *Functional Data Analysis*. Springer-Verlag.

Ramsay JO, Wickham H (2006). **fda***: Functional Data Analysis*. R package version 1.1.6, URL http://www.functionaldata.org/.

Rangarajan A, Chui H, Bookstein FL (1997). "The Softassign Procrustes Matching Algorithm." In JS Duncan, G Gindi (eds.), "Information Processing in Medical Imaging, 15th International Conference, Poultney, Vermont, USA, June 9–13, 1997, Proceedings," volume 1230 of *Lecture Notes in Computer Science*. Springer-Verlag.

Rangarajan A, Gold S, Mjolsness E (1996). "A Novel Optimizing Network Architecture with Applications." *Neural Computation*, **8**, 1041–1060.

Rangarajan A, Yuille A, Mjolsness E (1999). "Convergence Properties of the Softassign Quadratic Assignment Algorithm." *Neural Computation*, **11**, 1455–1474.

Reiner E, Abbey LE, Moran TF, Papamichalis P, Schafer RW (1979). "Characterization of Normal Human Cells by Pyrolysis Gas Chromatography Mass Spectrometry." *Biomedical Mass Spectrometry*, **6**, 491–498.

Sakoe H, Chiba S (1978). "Dynamic Programming Algorithm Optimization for Spoken Word Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **26**, 43–49.

Sinkhorn R (1996). "A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices." *Annals of Mathematical Statistics*, **35**, 876–879.

Steen H, Jebanathirajah JA, Springer M, Kirschner MW (2005). "Stable Isotope-free Relative and Absolute Quantitation of Protein Phosphorylation Stoichiometry by MS." *Proceedings of the National Academy of Sciences*, **102**, 3948–3953.

Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A, Le QT (2004). "Sample Classification from Protein Mass Spectrometry, by Peak Probability Contrasts." *Bioinformatics*, **20**(17), 3034–3044.

Tomasi G, van den Berg F, Andersson C (2004). "Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data." *Journal of Chemometrics*, **18**, 231–241.

van Nederkassel AM, Daszykowski M, Eilers PHC, Vander Heyden Y (2006). "A Comparison of Three Algorithms for Chromatograms Alignment." *Journal of Chromatography A*, **1118**, 199–210.

Walczak B, Wu W (2005). "Fuzzy Warping of Chromatograms." *Chemometrics and Intelligent Laboratory Systems*, **77**, 173–180.

Wang CP, Isenhour TL (1987). "Time-Warping Algorithm Applied to Chromatographic Peak Matching Gas Chromatography/Fourier Transform Infrared/MassSpectrometry." *Analytical Chemistry*, **59**, 649–654.

Affiliation:

Marc Kirchner, Benjamin Saussen, Fred A. Hamprecht
Multidimensional Image Processing
Interdisciplinary Center for Scientific Computing
University of Heidelberg
D-69120 Heidelberg, Germany
E-mail: Fred.Hamprecht@iwr.uni-heidelberg.de
URL: http://klimt.iwr.uni-heidelberg.de/mip/

Hanno Steen, Judith A. J. Steen, Fred A. Hamprecht
Proteomics Center at Children's Hospital Boston
Department of Pathology, Enders 1155
320 Longwood Ave
Boston, MA 02115, United States of America

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 18, Issue 4 *Submitted:* 2006-09-30
January 2007 *Accepted:* 2007-01-10