BIOINFORMATICS
Vol. 22 no. 14 2006, pages e132–e140
doi:10.1093/bioinformatics/btl219

Semi-supervised LC/MS alignment for differential proteomics

Bernd Fischer[1,*], Jonas Grossmann[2], Volker Roth[1], Wilhelm Gruissem[2], Sacha Baginsky[2] and Joachim M. Buhmann[1]

[1]Institute of Computational Science, ETH Zurich and [2]Institute of Plant Sciences, ETH Zurich, Switzerland

ABSTRACT

Motivation: Mass spectrometry (MS) combined with high-performance liquid chromatography (LC) has received considerable attention for high-throughput analysis of proteomes. Isotopic labeling techniques such as ICAT [5,6] have been successfully applied to derive differential quantitative information for two protein samples, however at the price of significantly increased complexity of the experimental setup. To overcome these limitations, we consider a *label-free* setting where correspondences between elements of two samples have to be established prior to the comparative analysis. The alignment between samples is achieved by nonlinear robust ridge regression. The correspondence estimates are guided in a semi-supervised fashion by prior information which is derived from sequenced tandem mass spectra.

Results: The semi-supervised method for finding correspondences was successfully applied to aligning highly complex protein samples, even if they exhibit large variations due to different biological conditions. A large-scale experiment clearly demonstrates that the proposed method bridges the gap between statistical data analysis and label-free quantitative differential proteomics.

Availability: The software will be available on the website http://people.inf.ethz.ch/befische/proteomics

Contact: bernd.fischer@inf.ethz.ch

1 INTRODUCTION AND RELATED WORK

A widely used approach to the sample-alignment problem fits a piece-wise linear function to maximize the correlation between the two samples. Methods of this kind are often characterized as *correlation optimized warping* (COW) (12). Other approaches are based on *hidden Markov models* (HMM) which formally define generative models for aligned samples, see e.g. Listgarten *et al*., (11). From a machine learning perspective, both COW and HMM methods are purely *unsupervised* in nature, since they do not exploit prior information of known correspondences. Both approaches share also the commonality that they have been solely applied to aligning *total ion counts*. Figure 1 depicts total ion count curves for two samples under two different biological conditions. Aligning these two samples is very difficult when the total ion counts are exclusively used as the information source.

In principle, both COW and HMM can be extended to aligning multi-dimensional data. It is, however, extremely difficult to handle LC/MS data of complex samples which are typically characterized by a very large input dimension (up to a mass range of 2500 Da for doubly charged peptides). The data analysis situation becomes even more complicated if we have to align highly heterogeneous samples that were taken under different biological conditions. Under these conditions one typically finds many peaks that do not match to *any* other peak in the second sample.

A first attempt to overcome these problems was made by Tibshirani *et al*. (14), who introduced an aligning technique based on hierarchical clustering.

In this paper we describe a new approach for LC/MS alignment exploiting additional information from sequenced tandem mass spectra rather than aligning only peaks from the LC/MS image. The second spectrometry stage is used to acquire sequence information. From a subset of these sequences which are identified in *both* samples, a time warping function is estimated by fitting a nonlinear regression function. Since there exists a number of false-identifications we use a *robust* regression model to reduce the sensitivity to outliers. Starting from an initial alignment hypothesis, we further improve the model by combining supervision information (sequenced peaks) and unlabeled information (all other peaks) within an iterative *self-training* scheme: the predictive variance is computed for each of the peaks, and peaks with a very small uncertainty are assigned a target value. Then, the model is re-trained based on the enlarged dataset, and the whole procedure is iterated until all peaks are labeled. This inclusion of unlabeled data yields an improved detection of peak correspondences. All free model parameters are selected by employing a cross-validation loop. With this novel machine learning technique we are able to align the underlying experiments of Figure 1.

2 EXPERIMENTAL SETTING AND DATA GENERATION

2.1 Liquid chromatography and mass spectrometry

Before analyzing the proteins in a cell, the proteins are digested by a specific enzyme like Trypsin, resulting in a mixture of small peptides. The peptides are separated by high-performance liquid chromatography. At (almost) equally spaced retention time steps a mass spectrum is acquired from the peptide sample eluting from the LC-column. The recording of a mass spectrum requires that a peptide is ionized and transferred into the gas phase, typically by electro-spray ionization. Most of the peptides are doubly or triply charged, but singly charged peptides also appear in proteomics experiments. The data are represented in form of a two dimensional

*To whom correspondence should be addressed.

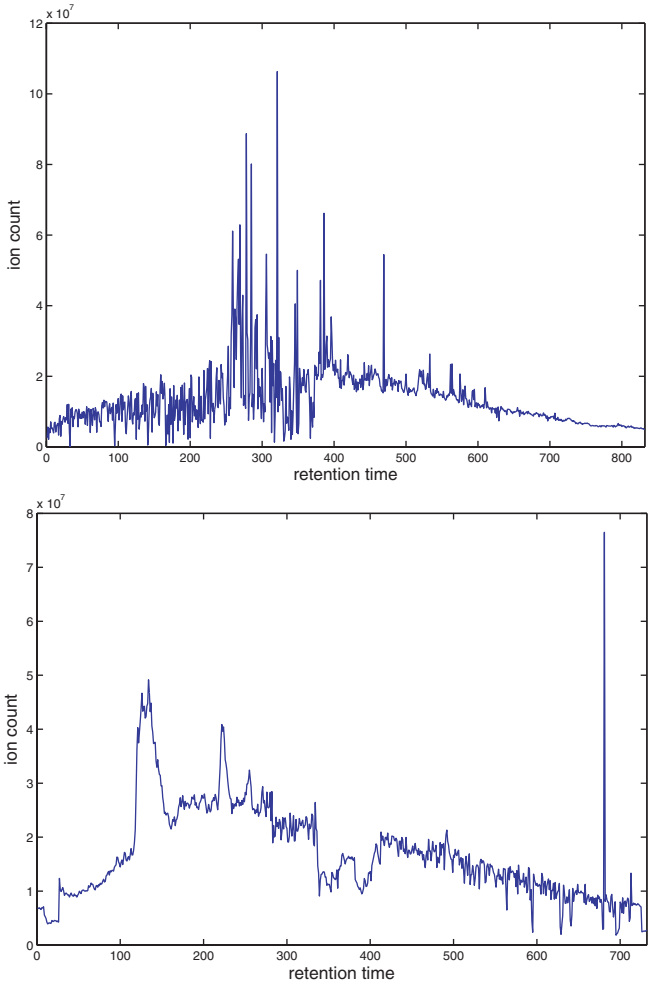

Fig. 1. The total ion count per time unit of two protein samples under two different biological conditions.

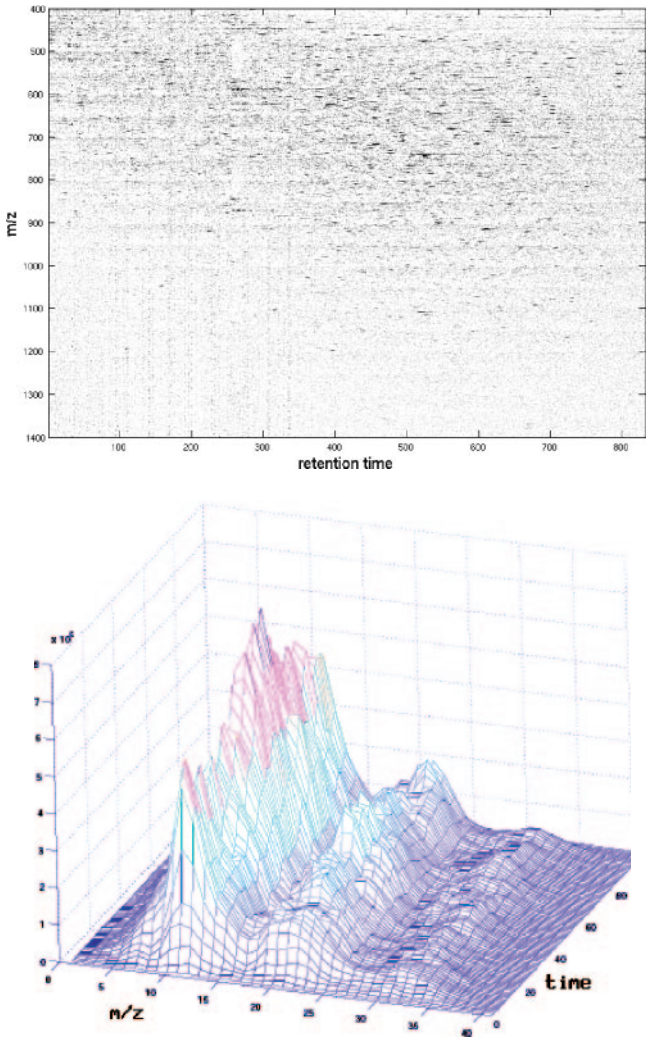

Fig. 2. Top: LC/MS image. The x-axis is the retention time, the y-axis is the peptide mass. Bottom: One peak in the LC/MS image accumulated over all singly charged peptides.

measurement, where one dimension is the retention time (t) and the other dimension is defined by the peptide mass over charge (m/z) (See Figure 2). We will refer to this two dimensional measurement as the LC/MS image. The local maxima in the LC/MS image correspond to different peptides with different m/z values. The bottom figure shows an accumulation of peaks over a large number of singly charged peptides. One can recognize three different isotopes for each peptide. Isotopes are common, since peptides are composed of a large amount of C-atoms. The integral over the peak area m_i yields the amount of ions of a specific peptide i.

2.2 Quantitative measurement

The over-all goal of quantitative proteomics is the estimation of the absolute protein expression. Let $I(p)$ denote the set of peptide indices for protein p. Assuming that all peptides $I(p)$ of a protein produce the same amount of ions and assuming a log-normal error distribution, one can estimate the log protein expression as

$$\widehat{\log e}_p = \frac{1}{|I(p)|} \sum_{i \in I(p)} \log m_i \,. \qquad (1)$$

The log-normal error model seems to describe expression levels well in practice, although we are not aware of any systematic study of this observation. The other assumption, however, that all peptides produce the same amount of ions rarely holds. The peak-area integrals are typically quite different for peptides of the same protein. One reason lies in the ionization efficiency of the peptides and suppression effects between peptides. An incomplete or overcomplete digestion process can also contribute to this discrepancy. There is still too little known about the reason for the different behavior of peptides. These uncertainties in the measurements render absolute quantitative proteomics infeasible today, but for peptide specific multiplicative errors, the ratio of peak area integrals can reliably be estimated. In our experience this assumption holds as long as the two samples are fairly similar. For two very different samples, peptide unspecific suppression effects play the major role.

Given two samples that both contain a certain peptide i and two corresponding measurements $m_i^{(1)}$ and $m_i^{(2)}$, the log-protein ratio in

both samples can be estimated by

$$\log r_p = \frac{1}{|I(p)|} \sum_{i \in I(p)} \log \frac{m_i^{(1)}}{m_i^{(2)}} . \quad (2)$$

A common procedure to measure peptides under two conditions is isotopic labeling like ICAT (5,6). The peptides in the two samples are marked with labels of different weights. The two samples are then mixed together and measured together. In the resulting LC/MS image, peptides of the two samples occur with a mass shift corresponding to the different weights of the labels. In addition to the fact that labels are still expensive, this approach carries the disadvantage that the two samples have to be mixed together. In many applications, however, it is advantageous to measure both samples separately. For example in cancer detection, one would like to first analyze a certain number of collected disease samples which then can be compared with patient probes without analyzing the disease samples over and over again.

Label-free techniques do not suffer from these shortcomings. Without the label information, on the other hand, one is forced to detect corresponding peaks in the two samples. In order to solve this correspondence problem we first have to shed some light on the procedure of *peak picking* which extracts peaks in the LC/MS image.

2.3 Peak detection

At the beginning of the analysis process, the mass spectrometry data is stored in a large data matrix, the columns of which represent mass spectra taken at different retention times. The *m/z* axis of these spectra is discretized in 1.00045 *Da* bins which can be justified as follows: If an amino acid is divided by its elementary mass (the number of protons and neutrons), the average mass of one elementary unit (a proton or neutron) is 1.00045 *Da*. Thus a peptide with 2000 elementary units has a mean mass of 2000.9 *Da*. The difference of 0.9 *Da* to the naively expected mean mass of 2000 *Da* is clearly detectable by our mass spectrometer and this mass correction significantly increases e.g. the peptide retrieval in de novo sequencing (3).

To ensure a standardized representation, each mass spectrum is normalized by its total ion count, i.e. by the sum over the spectrum. In the next step of the analysis process we measure the background noise level by median filtering over a window of ±50 in time and mass direction. This estimated noise level is then subtracted from the measurements. An entry in the LC/MS matrix is marked as a *peak area*, if the mean over ±5 in time direction and +1 in mass direction exceeds at least 3.0 times the mean over pixels surrounding the potential peak. The local maximum in each connected component defines the peak position with time and mass coordinates. Figure 3 shows the detected peaks in the LC/MS image.

2.4 Sequence identifications

At this stage of the analysis process the amino-acid sequence of the detected peaks is not available. We can, however, acquire sequence information for a certain fraction of peaks by way of *Tandem mass spectrometry*. From a measured MS spectrum a MS/MS device selects one of the peaks exceeding a predefined level. The ions in a small mass window around the selected mass are stabilized in an ion trap and fragmented by collision with a noble gas. The mass spectrum of the fragment ions contains information about the peptide sequence. The tandem mass spectra are denoted MS/MS spectra to distinguish them from standard MS spectra. Searching the spectrum against a database (2,7) produces hypotheses about the underlying peptide sequence. The hypothesized sequences are then validated by using *PeptideProphet* (10). In our experiments we consider spectrum identifications with a posterior probability $p \geq 0.97$ as being valid. Successful sequence identification without database knowledge is still a challenging problem. We have shown that small subsequences can be identified by *de novo* peptide sequencing (3) in many cases. In this work, however, we only use the database search results.

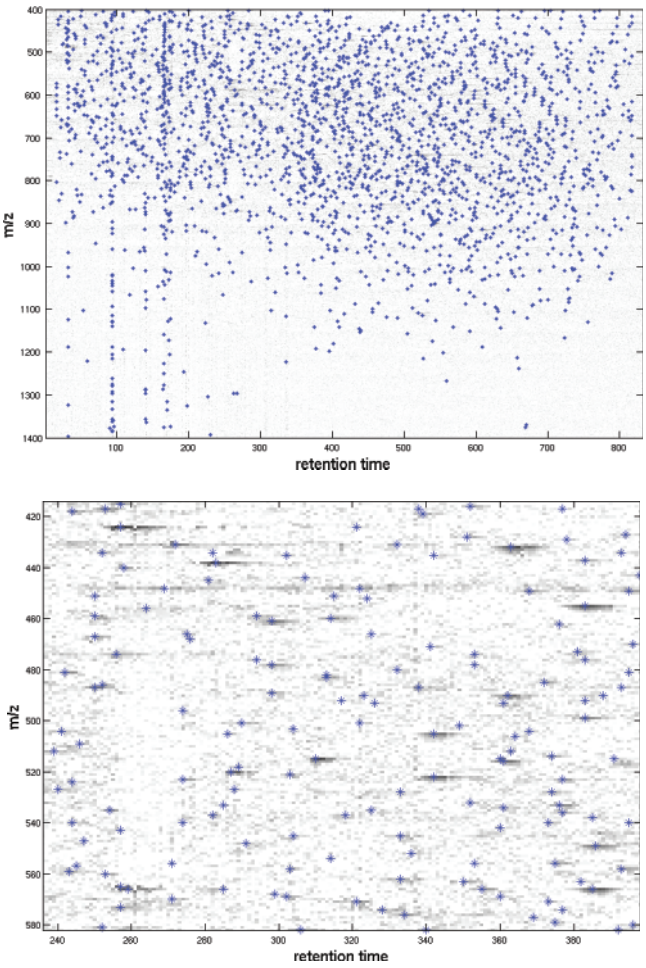

Fig. 3. Top: The detected peaks in the LC/MS image. Bottom: detailed view of sub-image.

To identify each MS/MS spectrum with one of the detected peaks in the LC/MS image, we search for a detected peak in the neighborhood of the mass/time coordinate of the MS/MS spectrum. We observed that in most cases the mass of the detected peak is correct or increased by 1 Da. Such increments might occur, if the first isotope is much larger than the mono-isotopic peak. Figure 4 depicts the fraction of sequenced MS/MS spectra that can be assigned to a peak. The quantity $w0$ denotes the size of the window, in which a peak is accepted if the mass is correct, and $w1$ is the corresponding window for mass differences of one. The asymmetry in the figure shows that the majority of peaks have the correct mass. Choosing

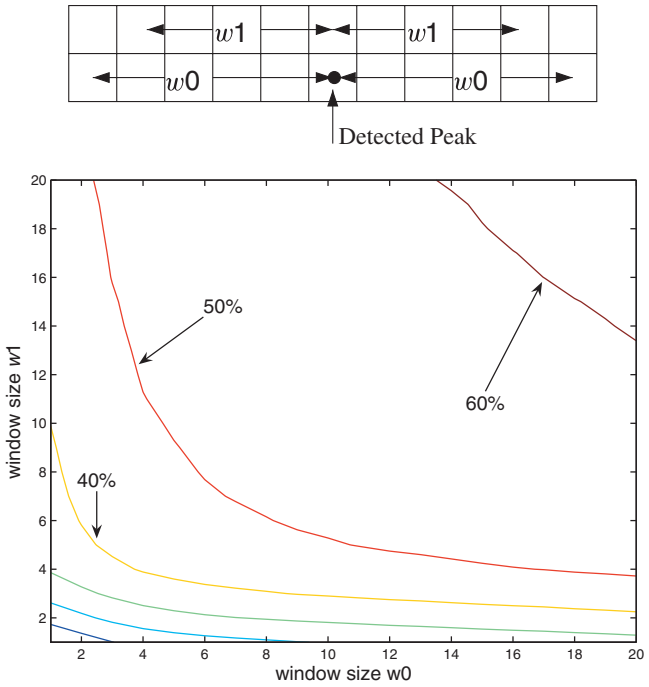

Fig. 4. Top: The window in which an identified spectrum is assigned to a peak. Bottom: The fraction of identified spectra that can be assigned to a peak for varying window sizes. Contour lines are spaced in 10% intervals.

$w0 = 10$ and $w1 = 5$ we can assign 52.2% of all identified sequences to a peak. This rate might be increased by using larger windows, however at the price of a higher false-positive rate. We will discuss this issue in more detail later. The large fraction of not assignable sequenced spectra is due to peaks that can hardly be distinguished from the background.

The search includes all singly, doubly, and triply charged peptides. Denoting the mono-isotopic mass of a peptide by m, the m/z-values of a singly ($i = 1$), doubly ($i = 2$) and triply ($i = 3$) charged peptide are observed as mass/charge ratios

$$\frac{m^{(i)}}{z} = \frac{m+i}{i} \tag{3}$$

due to protone capture. On average there are about 5000 peaks per LC/MS image from which roughly 200 could be sequenced. Figure 5 depicts the distribution of the different charge states over the LC/MS image. The green circles show the singly charged peptides, the red crosses are the doubly charged peptides and the blue filled circles are the triply charged ones.

2.5 Scenarios in quantitative proteomics

The analysis process described above extracts two different types of information from the mass spectrometry data:

- a list of peaks in the LC/MS image, and
- sequence information for a small subset of the peak list.

A quantitative analysis based on these input data can pursue different goals:

(i) in a **classification scenario** one would like to separate a certain protein sample under one biological condition (extracted e.g. from a diseased patient) from samples under another biological condition (extracted e.g. from a control group). For the mere classification task one does not need the peptide sequence information. One rather tries to find as many corresponding peaks between two samples of different biological conditions as possible.

(ii) A different scenario is known as **biomarker discovery** (9). In addition to classification, one would like to identify proteins or peptides which are causally related to a certain biological condition (e.g. a certain disease). From a machine learning point of view this identification problem defines a *feature selection* task. Having selected 'relevant' features one is typically interested in the underlying sequences. Thus, if we pursue biomarker discovery as our goal, we have to *sequence* as many peptides as possible. Ultimately, we try to compare the complete proteome using these processing steps.

In this paper we will show that the number of peak correspondences (for classification) as well as the number of sequence identified correspondences (for biomarker discovery) can be increased by combining labeled and unlabeled information.

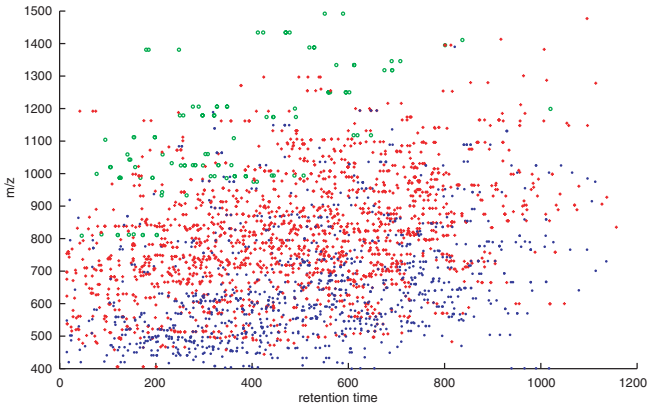

Fig. 5. Distribution of charge values over LC/MS image. Green circles: singly charged peptides. Red crosses: Doubly charged. Blue filled circles: Triply charged.

2.6 Sample preparation

The peptides we used for the analysis were derived from plant cell culture samples that were exposed to different illumination programs (light versus dark). The proteins are fractionated by SDS-PAGE and in-gel digested. The peptide mixture was loaded onto a C18 reversed phase column and eluted with a gradient developed from solvent A (5% ACN, 0.2% formic acid) and solvent B (80% ACN, 0.2% formic acid). Gradient shape was as follows: 26 minutes 100% solvent A, within 0.2 minutes up to 5% solvent B, within additional 69 minutes up to 55% solvent B and in one additional minute up to 100% B. The flow rate at the tip of the column was adjusted to ≈200 nl/min. The chromatography (LC) was coupled online to an LTQ ion trap mass spectrometer (Thermo-Finnigan, San Jose, CA, USA) equipped with a nanospray ionization source. Mass analysis was performed with a spray voltage of 2.0–2.5 kV and one MS full scan followed by three data-dependent MS/MS scans of the three most intensive parent ions. The dynamic exclusion function was enabled to permit one measurement of a particular

parent ion followed by an exclusion of the aquisition of MS/MS spectra for this parent ion over a periode of 4 min.

3 LC/MS ALIGNMENT

When comparing two subsequent LC/MS scans, slight changes of the time scale can often be observed in different experimental situations. To compensate these time differences, an alignment function of the form $f : t^{(1)} \mapsto t^{(2)}$ maps the time scale of one experiment to that of the second experiment. Instead of directly mapping the scale itself, one can alternatively map one scale to the *scale differences* between the two samples:

$$g : t^{(1)} \mapsto t^{(2)} - t^{(1)}. \tag{4}$$

This formulation provides a clear visualization of the inherent non-linearities of the warping process. Within the subset of peaks sequenced in the second MS stage, we typically find an overlap of 10-70 identified peaks that are common in both experiments. Figure 6 depicts such time-warping functions learned from the subset of common peptides for two different pairs of biological samples. The non-linear relationship between the time-scales is clearly visible in the top panel.

3.1 Warping by way of robust regression

Identifying the $t_i^{(1)}$-values with x_i, and the time differences $t_i^{(2)} - t_i^{(1)}$ with y_i, the warping function depicted in Figure 6 is determined by first expanding the x-values in a k-th order polynomial basis

$$\phi_i := \phi(x_i) = (1, x_i, x_i^2, \ldots, x_i^k)^t, \tag{5}$$

and then by fitting a robust ridge-regression model. The latter finds the $k+1$ dimensional weight vector β which minimizes

$$\sum_{i=1}^{n} L_c(\phi_i^t \beta - y_i) + \lambda \beta^t \beta, \tag{6}$$

where $L_c(\xi)$denotes a robust loss function of Huber's type:

$$L_c(\xi) = \begin{cases} c\,|\,\xi\,| - \dfrac{c^2}{2}, & \text{for}\,|\,\xi\,| > c \\ \dfrac{\xi^2}{2}, & \text{for}\,|\,\xi\,| \le c. \end{cases} \tag{7}$$

Both the degree k of the polynomial and the ridge-penalty λ are chosen by 10-fold cross-validation. The reader should notice that the above nonlinear regression model is equivalent to using a *kernel regression model* with polynomial kernal of degree k. For computational reasons, in this special application it is better to *explicitly* expand the input data in the polynomial basis, rather than using the kernelized version.

In the usual regression setting, the observations y are assumed to be generated by corrupting the values of $f(x_i) = \phi_i^t \beta$ by additive noise that follows some density $p(\xi)$. Huber's loss function turns out to be optimal (in the sense that it guarantees the smallest loss in a worst case scenario), if the true noise density is a mixture of two components, one of which is known to be Gaussian distributed and the other one is an arbitrary density (8). Huber's loss function penalizes large deviations $|\,\xi\,| > c$ only linearly. Thus, it is superior to its standard quadratic counterpart in situations where the data contains outliers which are generated by an unknown and possibly

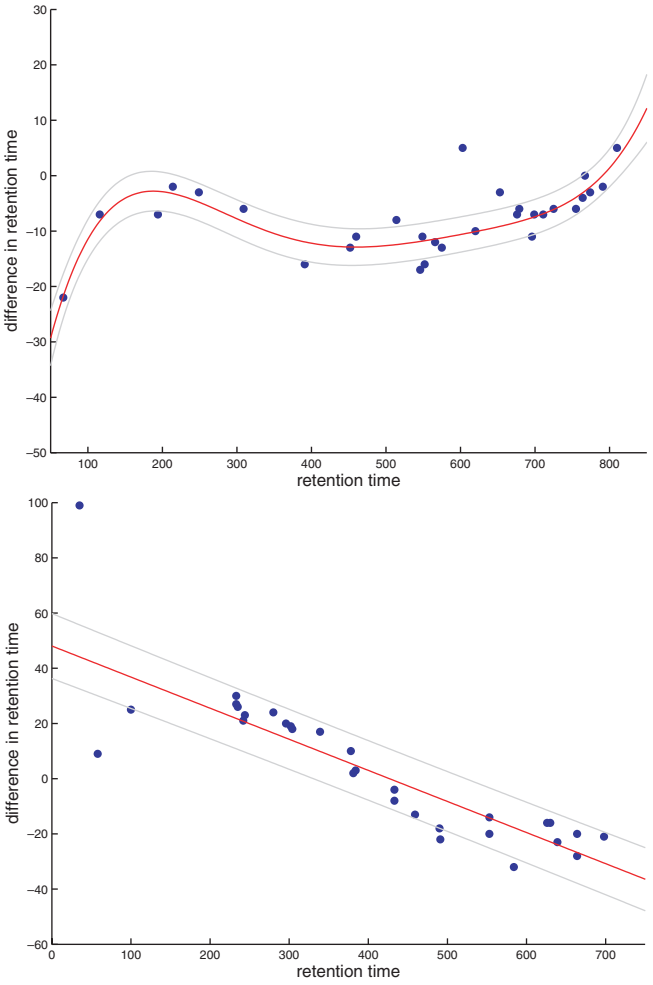

Fig. 6. Examples of two different alignments. On the x-axis the retention time is plotted, on the y-axis the difference in retention time. The red curve depicts the estimated warping function, the light gray ones show 1σ-confidence intervals.

highly fluctuating noise source. The parameter c is typically estimated from the data in an iterative fashion as a multiple of the standard deviation of the observed residuals. A common scaling formula is $c = 1.345\sigma$, which yields 95% efficiency when the errors are normal, and still protects against outliers. Usually a robust measure of spread is employed in preference to the standard deviation of the residuals. For example, a common approach is to choose $\hat{\sigma} = MAR/0.6745$, where MAR is the median absolute residual. This choice defines an unbiased estimator of the standard deviation for Gaussian data, see (4).

The optimal weight vector β that minimizes eq. (6) is found iteratively as the solution of a re-weighted least squares problem:

$$\beta^{\text{new}} = [\Phi^t \Omega(\beta) \Phi + 2\lambda I]^{-1} \Phi^t \Omega(\beta) y, \tag{8}$$

where Φ denotes the (transformed) data matrix with rows ϕ_i, and $\Omega(\beta)$ denotes the diagonal matrix

$$\Omega(\beta) = \text{diag}\{\omega([\Phi\beta - y]_i)\}, \tag{9}$$

with $\omega(\xi) := (1/\xi) \cdot \frac{\partial L_\varepsilon(\xi)}{\partial \xi}$. The final entries Ω_{ii} define weights for the individual training data $\phi(x)_i$.

3.2 Semi-supervised alignment

In the above derivation, the regression function is learned exclusively from the subset of identified correspondences in both samples. Due to technical limitations, the number of MS/MS spectra and thus the number of peptide sequence identifications is usually relatively small. We will now exploit the ideas of *self-training* (13), to additionally extract the information contained in the remaining peaks. Self-training is an incremental algorithm that labels the unlabeled data and converts the most confidently *predicted* data points into labeled training examples. This iteration proceeds until all the unlabeled data are consistently labeled. In order to apply this mechanism to our LC/MS alignment problem, we have to derive a formula for the *predictive uncertainty* of test data.

We denote by Φ_G the subset of training data which have been assigned a weight $\Omega_{ii} = 1$ in the robust regression procedure defined in eq. (8). These data points have small residuals $|\xi| \leq k$ which are penalized quadratically by the robust loss function eq. (7). Thus, for these points the *Gaussian* noise assumption is valid. Since in this case the posterior distribution is also Gaussian, a Bayesian treatment of regression allows us to derive an analytical expression for the uncertainty of the prediction for a new data point x_*:

$$\begin{aligned} Var[f(x_*)] &= E_{\beta|X}[(f(x_*) - E[f(x_*)])^2] \\ &= \sigma^2 \phi^t(x_*)(\lambda I + \Phi_G^t \Phi_G)^{-1} \phi(x_*). \end{aligned} \tag{10}$$

The total predictive variance, $Var[y(x_*)]$, is the sum of the noise variance σ^2 and the variance about the mean, $Var[f(x_*)]$, since both sources of variation are uncorrelated, see e.g. (1) for details. For estimating the noise variance one might again use the above equation $\hat{\sigma} = MAR/0.6745$ applied to the data in Φ_G.

Our adaption of the self-training method now proceeds as follows:

Initialize: train the model on the correspondences verified by sequencing.

Iterate:

(i) for a peak which elutes at time $t_i^{(1)}$ in the first LC/MS image, predict the time difference $t_i^{(2)} - t_i^{(1)}$;

(ii) for every such predicted peak, compute its predictive variance;

(iii) for the 10% most certain peaks, search for a corresponding peak in the second LC/MS image within a certain window.

(iv) include all found correspondences into the training set, and retrain the model;

Until: No more peaks are found within a 2σ-confidence interval around the current fit.

Figure 7 shows the outcome of this semi-supervised learning algorithm for the two samples that were analyzed previously in Figure 6. The labeled objects are colored dark blue. Compared to the alignment computed exclusively on the labeled objects (cf. Figure 6), the inclusion of unlabeled objects makes it possible to model more details of the warping function. Compared to the supervised solutions, where often only a straight line can be reliably fitted to the data, the semi-supervised solutions typically use regression models of higher complexity (measured in terms of the polynomial degree k in the expansion eq. (5), which is automatically selected by cross validation).

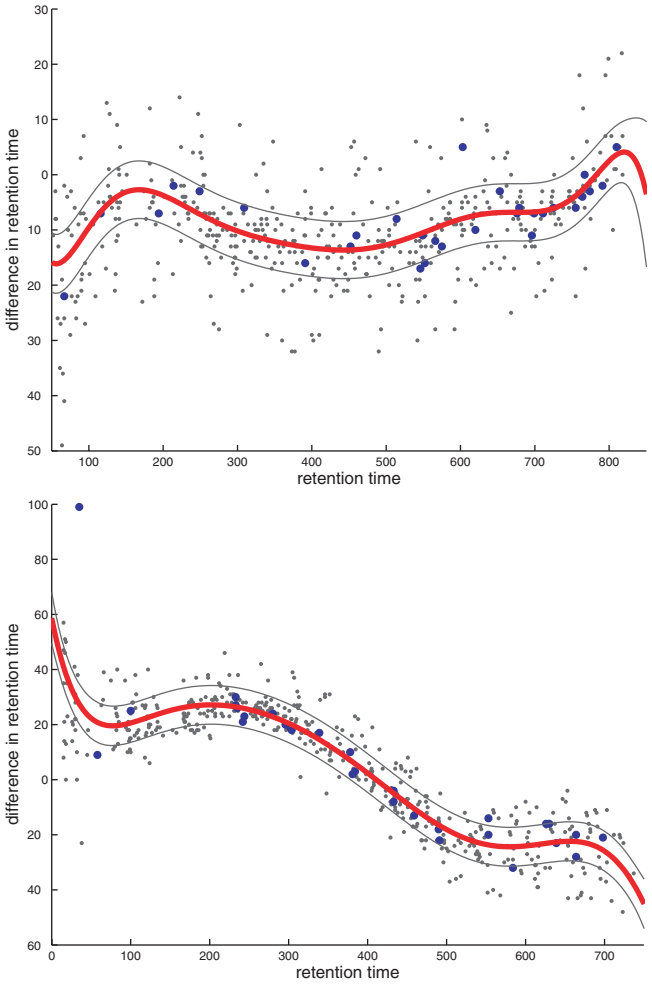

Fig. 7. Example of two different alignments for semi-supervised learning. On the x-axis the retention time is plotted, on the y-axis the difference in retention time. The blue (light gray) dots are the sequenced (non-sequenced) peaks. The light gray curves depict 1σ-intervals of the predictive uncertainty.

3.3 Detecting peak correspondences

First we analyze the performance of the alignment in the **classification** scenario, where all peaks (sequenced as well as unsequenced) are aligned. The alignment function computed by minimizing eq. (6) treats the two samples in an non-symmetrical fashion, since it warps the first time scale to the second. In order to derive symmetric correspondences between peaks, we predict the retention times in both directions separately, which allows us to easily check the self-consistency of the prediction model. Given a peak in sample A, our method predicts the retention time in sample B. If we have detected a peak in sample B within a window w around the predicted peak position, we denote this a (directed) correspondence. Here again we tolerate a mass difference of at most ±1 Da. Predicting retention time in both directions between sample A and sample B gives us a list of (directed) correspondences from sample A to sample B and a list of (directed) correspondences

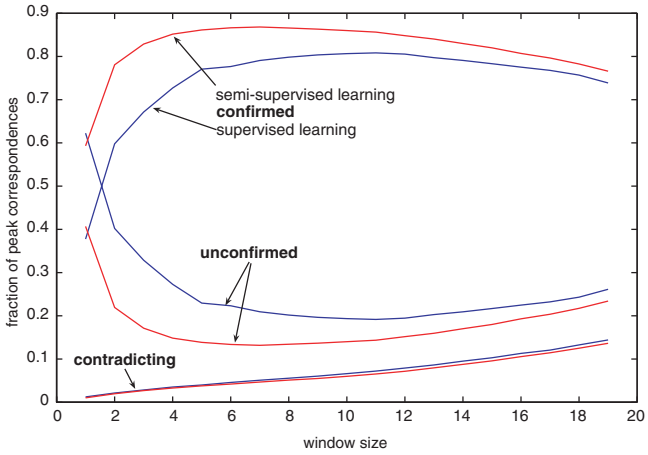

Fig. 8. Fractions of confirmed/unconfirmed/contradicting peak correspondences for the semi-supervised model (red) and the purely supervised model (blue) as a function of window size.

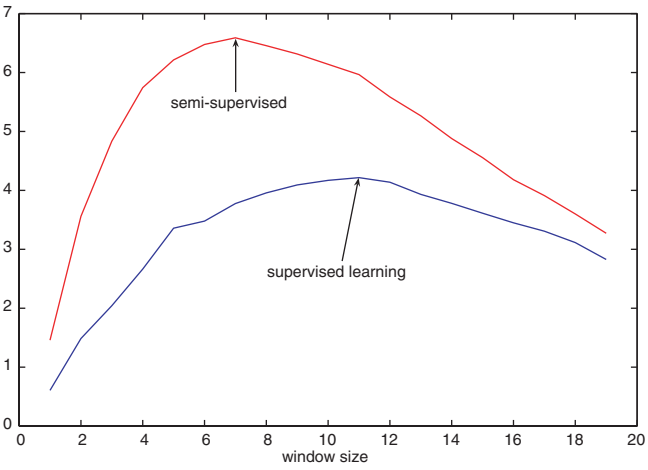

Fig. 9. Performance ratio #(confirmed)/[#(unconfirmed) + #(contradicting)] correspondences for the semi-supervised model (red) and the supervised model (blue).

from sample *B* to sample *A*. A correspondence is called *confirmed* if we find a correspondence in both directions. If we find a peak only in one of the directions, we call the correspondence *unconfirmed*. If we obtain two different mappings for one peak, we declare the 'correspondence' as *contradicting*. Denoting by n_1 the number of confirmed, by n_2 the number of unconfirmed and by n_3 the number of contradicting correspondences, the respective rates $n_i/(n_1 + n_2 + n_3)$ are depicted in Figure 8. It is obvious that the fraction of contradicting correspondences monotonically increases if the window is enlarged. For very small windows most correspondences remain unconfirmed, whereas the fraction of confirmed correspondences attains a maximum for windows of intermediate size. In practice, we have to balance the number of confirmed correspondences against the unconfirmed and/or contradicting ones. Figure 9 shows the quotient $n_1/(n_2 + n_3)$ both for the semi-supervised and supervised

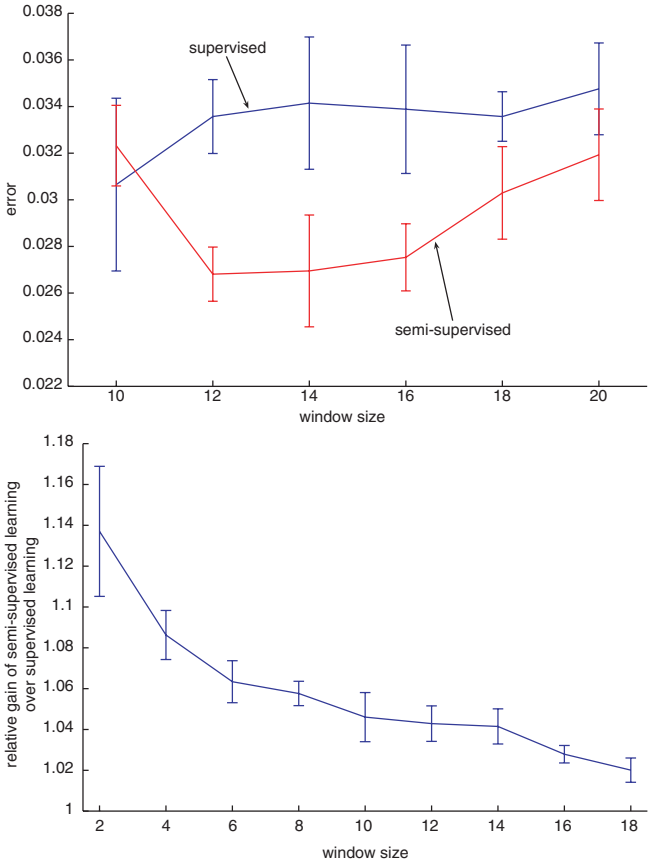

Fig. 10. Top: cross-validation error of the alignment. Bottom: The gain of semi-supervised learning over supervised learning.

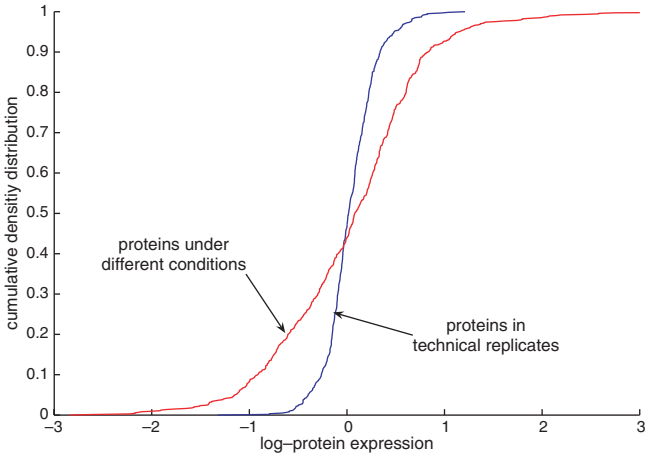

Fig. 11. Cumulative distribution function of the protein ratios for replicate measurements and for different conditioned samples.

variants. These two curves nicely summarize the benefits of the inclusion of unlabeled data: the maximum is higher (which is obviously desirable), and it is attained at smaller window sizes, which is also desirable, since it yields better localization in the mass-retention time space.

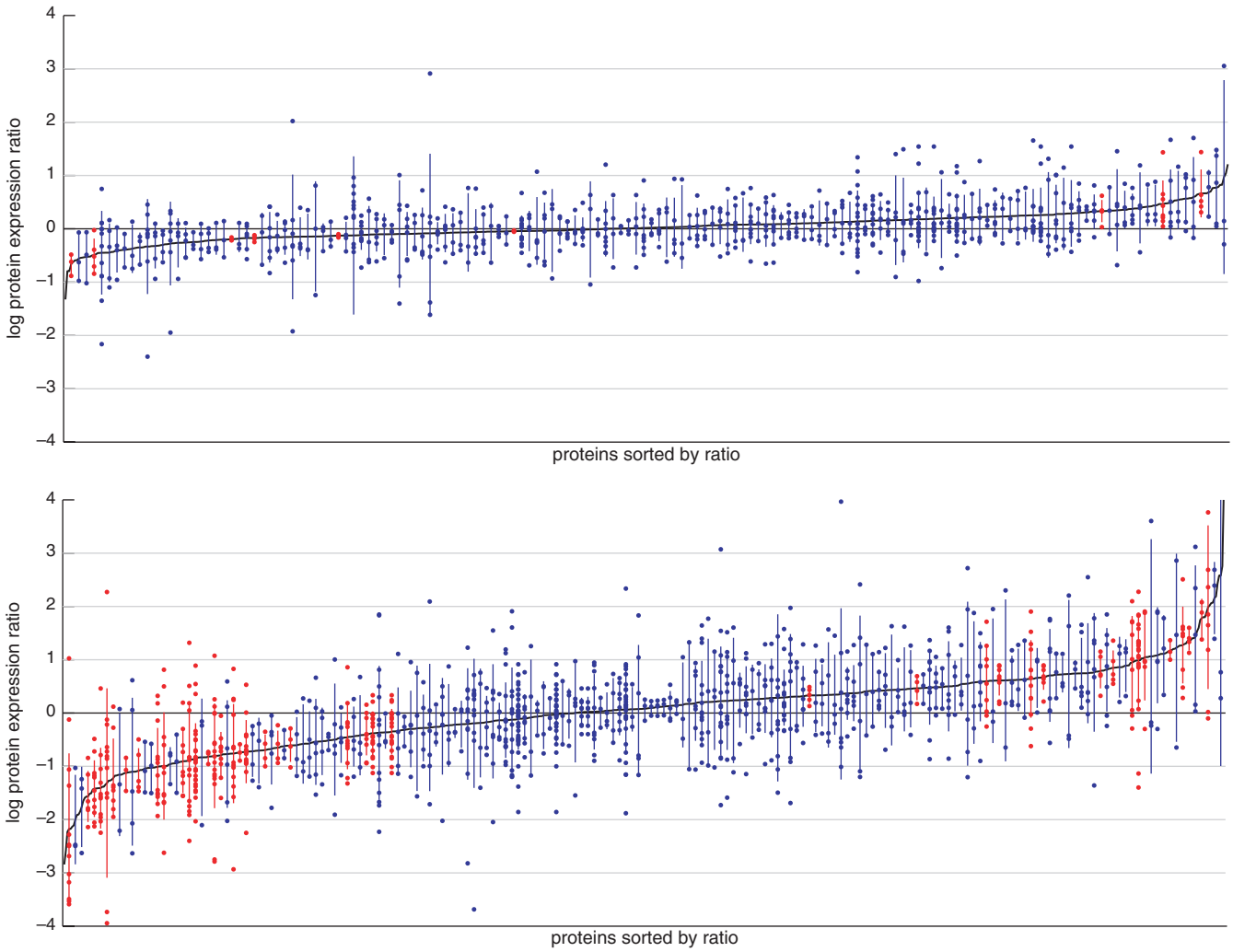

Fig. 12. Top Protein ratios for replicate measurement. Bottom: Protein ratios for different biologically conditioned samples.

3.4 Cross-validation

To show the efficiency of our approach we test it by cross-validation for **biomarker discovery**. For each alignment we divide the set of the known correspondences in a training set and a test set with size proportions (75%/25%). Only the training set is used as supervision information during alignment. On the test set we evaluate the error of the alignment. Such an error occurs, if the known sequences are assigned to different peaks. Figure 10 (top) depicts the cross-validation error for sequenced peaks. The error is plotted against the window size of acceptance for peak correspondences. An error of less than 0.03 is achieved for window sizes smaller than 15. Window sizes smaller than 10 are excluded from the plot, because the corresponding error bars are extremely large, since only very few identifications could be found. Compared to the fraction of contradicting peaks in Figure 8, the error rate on the subset of sequenced peaks is much smaller. The reason for this reduced error is that many of the contradicting peaks in the *un*supervised setting are not counting for an error in this supervised setting: a contradiction in the unsupervised setting occurs, if two different peaks from sample 1 are assigned to the same peak in sample 2. In the supervised setting such an inconsistency produces an error only if both peaks from sample 1 are *differently sequenced*. Sometimes the peak picking algorithm finds two peaks where only one peak should be placed. The sequenced MS/MS spectrum, however, is only assigned to *one* of the two ''pseudo''-peaks. In the unsupervised setting, such a situation would be treated as a contradiction, whereas in the supervised setting no error occurs.

On the bottom of the figure the gain of the semi-supervised method is plotted. We defined the gain as the ratio of confirmed correspondences for semi-supervised learning compared to supervised learning. One achieves 5% more assignments with semi-supervised learning than supervised learning at a window size of 15. Here again the improvement due to the semi-supervised method increases with smaller window size.

4 DIFFERENTIAL PROTEIN EXPRESSION

The first step towards biomarker discovery requires to compute a list of differential protein expression values. To increase the number of

sequenced peptides in each LC/MS image, we generate three replicate LC/MS/MS measurements per condition. To compare two differently conditioned samples, we compute pairwise alignments of all six LC/MS images and predict the retention time of the peptides that have been sequences from each LC/MS image to all others. This procedure yields an extensive increment in the number of sequenced peptides in each single LC/MS image. For each protein we obtain a collection of differential peptide measurements, from which the log protein ratio is estimated according to eq. (2).

To demonstrate the possibility to derive differential quantitative measurements from biological samples, we estimate the expression ratio both for replicate measurements and for differently conditioned samples. Figure 12 shows the differential protein expression. For better visualization, only a (randomly drawn) subsample of the proteins is plotted. Each protein corresponds to one column. The dots on the columns depict the differential peptide measurements. The vertical lines indicate one standard deviation. A t-test with a significance level of 0.03 rates 3.9% (24 out of 610) of the peptides as significantly over- or underexpressed for the replicate measurements. The significantly over-/under expressed proteins are colored red. For the biologically different samples (bottom panel) one can detect 24.5% (165 out of 735) of the proteins as significantly under- or overexpressed. These six times higher rate of significantly different expression levels between biologically different samples and technical replicates demonstrate that our statistical analysis is sensitive to changes in conditions. We conclude that we are able to recognize differences in protein expression by label-free differential quantitative proteomics. To conclude that the differences are caused by the different conditions, one should still compare the result with biological replicates.

5 DISCUSSION AND CONCLUSION

In the recent years the use of LC/MS measurements has received considerable attention for high-throughput analysis of proteomes. For *quantitative* differential measurements it is commonly accepted that isotopic labeling techniques such as ICAT are needed for a reliable quantitative comparison of two protein samples. These labeling techniques are not ideal, however, because they require a significantly increased complexity of the experimental setup and the necessity to mix the two labeled samples from different biological conditions. The latter is particularly problematic in applications like *biomarker discovery* where one would like to treat samples from different biological conditions separately in order to avoid a time-consuming and costly re-analysis of the, e.g., disease-specific reference sample.

As an alternative approach, we consider a *label-free* setting for comparative proteomics. The absence of isotopic labels that could guide the search for correspondences, however, imposes a severe *alignment problem* between the elements of the two samples from different biological conditions. Current approaches to solve this problem try to find alignments solely on the basis of the observed LC/MS measurements while ignoring potentially relevant additional information from the underlying sequences. In contrast to these approaches, we propose to use *tandem mass spectrometry* to extract partial sequence information of the peptides contained in the samples. Based on this subset of sequenced peptides, we compute a ''seed'' alignment by estimating a *nonlinear robust regression* function which warps one time scale into the other. Within a *semi-supervised* learning framework, this seed alignment is iteratively refined by successively including the mass peaks for which no sequence information is available. By assessing the self-consistency of the time warping in both directions, we have shown that this refinement process significantly improves the quality of the alignment.

In a large-scale experiment we have demonstrated that our method is capable of aligning *highly complex* protein samples, even if they exhibit *large variations* due to different biological conditions. It is possible to *reliably discriminate* between technical replicates and truly different biological conditions. We conclude that the proposed method bridges the gap between statistical data analysis and label-free quantitative differential proteomics.

ACKNOWLEDGEMENTS

This research was supported by the Functional Genomics Center Zurich and it was partially funded by the Swiss Initiative in Systems Biology (SystemsX: CC-SPMD and C-MOP) and by ETH-grants TH-5/04-3 and TH-41/02-2.

REFERENCES

[1] Box,G.E.P. and Tiao,G.C. *Bayesian Inference in Statistical Analysis*. Wiley, New York, 1992.

[2] Eng,J.K., McCormack,A.L. and J.R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Am. Soc. for Mass Spectrometry*, **5**(11), 976–989, 1994.

[3] Fischer,B., Roth,V., Roos,F., Grossmann,J., Baginsky,S., Widmayer,P., Gruissem,W., and Buhmann,J.M. NovoHMM: A hidden markov model for de novo peptide sequencing. *Anal. Chem.*, **77**(22), 7265–7273, 2005.

[4] Fox,J. *Applied Regression, Linear Models, and Related Methods*. Sage, 1997.

[5] Gygi,S.P., Rist,B., Gerber,S.A., Turecek,F., Gelb,M.H. and Aebersold,R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotech.*, **17**, 994–999, 1999.

[6] Han,D.K., Eng,J., Zhou,H. and Aebersold,R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotech.*, **19**, 946–951, 2001.

[7] Hirosawa,M., Hoshida,M., Ishikawa,M. and Toya,T. Masqot: multiple alignment system for protein sequences based on three-way dynamic programming. *Comp. App. in the Bioscience*, **9**(2), 161–167, 1993.

[8] Huber,P.J. *Robust Statistics*. Wiley, New York, 1981.

[9] Jacobs,J.M., Adkins,J.N., Qian,W.-J., Liu,T., Shen,Y., D.G. Camp II, and Smith,R.D. Utilizing human blood plasma for proteomic biomarker discovery. *J. of Proteome Res.*, **4**, 1073–1085, 2005.

[10] Keller,A., Nesvizhskii,A.I., Kolker,E. and Aebersold,R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392, 2002.

[11] Listgarten,J., Neal,R.M., Roweis,S.T. and Emili,A. Multiple alignment of continuous time series. In *NIPS 17*, pages 817–824, 2005.

[12] Vest Nielsen,N.-P., Carstensen,J.M. and Smedsgaard,J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. of Chromatography A*, **805**, 17–35, 1998.

[13] Nigam,K. and Ghani,R. Analyzing the effectiveness and applicability of co-training. In *CIKM '00*, pages 86–93, 2000.

[14] Tibshirani,R., Hastie,T., Narasimhan,B., Soltys,S., Shi,G., Koong,A. and Le,Q.-T. Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, **20**, 3034–3044, 2004.