

# A Fast and Accurate Algorithm for the Quantification of Peptides from Mass Spectrometry Data

Ole Schulz-Trieglaff<sup>1,2,4</sup>, Rene Hussong<sup>3</sup>, Clemens Gröpl<sup>2</sup>,  
Andreas Hildebrandt<sup>3</sup>, and Knut Reinert<sup>2</sup>

<sup>1</sup> Max Planck Research School, Berlin, Germany

<sup>2</sup> Department of Computer Science and Mathematics, Free University Berlin

<sup>3</sup> Center for Bioinformatics, Saarland University

<sup>4</sup> [trieglaof@inf.fu-berlin.de](mailto:trieglaof@inf.fu-berlin.de)

**Abstract.** Liquid chromatography combined with mass spectrometry (LC-MS) has become the prevalent technology in high-throughput proteomics research. One of the aims of this discipline is to obtain accurate quantitative information about all proteins and peptides in a biological sample. Due to size and complexity of the data generated in these experiments, this problem remains a challenging task requiring sophisticated and efficient computational tools.

We propose an algorithm that can quantify even low abundance peptides from LC-MS data. Our approach is flexible and can be applied to preprocessed and raw instrument data. It is based on a combination of the sweep line paradigm with a novel wavelet function tailored to detect isotopic patterns. We evaluate our technique on several data sets of varying complexity and show that we are able to rapidly quantify peptides with high accuracy in a sound algorithmic framework.

## 1 Introduction

Quantitative proteomics is increasingly developing into one of the cornerstones of fundamental research in the life sciences and of clinical studies [17, 18, 21]. In a typical experimental setting, the protein sample is subjected to a proteolytic digestion yielding a mixture of peptides which is inserted into a chromatographic column for a first separation. The peptides elute at different retention times due to their interaction with the stationary phase of the column. The protein digest is thus separated according to a physical property, like the peptide’s hydrophobicity in the case of reversed-phase (RP) liquid chromatography. The fractions of the analyte are transferred into a mass spectrometer where they are ionized and separated by their mass/charge ratio.

The resulting data consists of a sequence of MS spectra (*scans*). Each scan gives a snapshot of the peptides eluting from the column during a fixed time interval. It consists of ion counts or *intensities* measured by the mass spectrometer within a certain interval of mass/charge ratios. The scans are acquired at (more or less equally spaced) periodic time intervals, and the collection of scans constitutes what we will call an *LC-MS map*.

**Problem statement.** Many applications in proteomics, such as mass-spectrometry based diagnostics, rely on an accurate, and – given the size of the data – fast quantification of proteins or peptides contained in a biological sample. Indeed, the quantification problem lies at the very base of a whole proteomics pipeline, where all subsequent steps depend on the quality of the data generated in the beginning. In this work, we report on a fast and reliable approach to peptide quantification tailored for very large LC-MS maps.

To accurately estimate the abundance of peptides in a biological sample, we need to collect all data points that are caused by a charge variant of a peptide, that is by all ions with identical sequence and charge. We will refer to this problem as the *peptide quantification problem*. The

detection of peptidic features in LC-MS spectra can be considerably improved by exploiting prior knowledge about the data produced during the experimental process: data points in an LC-MS map belonging to the same peptide are locally highly correlated in the mass over charge dimension as well as in the retention time domain. The atoms contained in a peptide occur in different isotopic variants, and the distribution of the naturally occurring isotopes gives rise to a characteristic *isotopic pattern* of adjacent peaks in the mass spectrum. Similarly, each peptide elutes over a certain interval of time from the column and can be observed in several consecutive scans. The *elution profile* ideally follows a normal distribution around its centroid, but fronting and tailing effects are frequently observed in practice [3].

In addition to these local relationships, non-local correlations can also be observed: different charge states of ionized peptides show up at distant  $m/z$  values, and different peptides originating from the same protein have distant RT (and  $m/z$ ) values. For the remainder of this work, however, we will restrict ourselves to local correlation effects.

**Previous work.** Several computational approaches to peptide quantification have been developed recently. Some of them are embedded into a software framework comprising other processing steps such as identification of proteins or alignment of LC-MS maps as well. A recent review of these software tools is given in [22], while [16] gives a more general overview of the computational problems in proteomics data analysis.

An algorithm for peptide quantification needs to address the following problems: in a first step, prominent data points (or *seeds*) in the LC-MS map need to be found. These seeds are data points that are very likely to be in the region (in  $m/z$  and retention time dimension) of data that can be attributed to a peptide charge variant. Second, we want to *extend* these seeding points to *regions* of interest in the LC-MS map. Due to posttranslational modifications, isotopic variants and different charge states of the same peptide, it is not feasible to restrict the search to single points in the spectrum. Rather, we always need to consider clusters of data points centered around the seeds. In the literature, several approaches have been proposed to identify such regions based on the intensity of the data points [1, 6, 15, 14, 27] or using image segmentation techniques [12, 26].

The image-based approach usually starts by *resampling* the data to obtain a gray-scale image from the LC-MS map [12, 26]. But the dimensions of LC-MS separation have quite different characteristics and require different handling, and a resampling diminishes the resolution of the data and will almost certainly lead to a loss of information.

If the set of candidate regions is chosen based on the intensity of single data points or local maxima in the LC-MS map, it is likely to contain many false positives, i.e. groups of data points caused by noise or contaminants of the sample and not by peptides. In addition, these approaches are hampered by the low signal-to-noise ratio of peaks at the beginning and end of the peptide elution period. Sophisticated methods are required to estimate the background noise in the spectra and to exclude outlier data points from the isotopic pattern before a peptide abundance can be estimated [14, 27]. In some cases, information from tandem spectra and database search is taken into consideration to increase the confidence in detected seeds [5, 14]. But due to the high error rates of current peptide identification algorithms [8], this approach has its own disadvantages.

After the extension one can employ an additional *refinement step* during which a theoretical peptide model is adjusted to the selected data points [1, 6, 12, 15]. The quality of this fit is taken as a measure of confidence that this region is indeed caused by a peptide. Regions with poor correspondence to the theoretical model are discarded. By summing the intensities of all data

points in the regions identified, we can obtain an estimate of the peptide that can be used for a relative quantification. This sequence of finding seeds, extension and refinement is a general concept, and the majority of algorithms follow these steps e.g. [1, 6, 12, 15].

A typical proteomic sample consists of several thousand peptides and clinical studies typically consist of hundreds of samples. It is therefore desirable to quantify all peptides in a sample as quickly as possible. The refinement step in particular takes considerable time and sometimes even requires a manual validation of the peptide candidates. An efficient algorithm that is suitable for real-world applications should thus aim for a low number of seeds. However, if seeds or regions in the map are chosen based on their intensity alone, we will obtain a large number of seeds, many of which will simply be caused by chemical noise or contaminants of the sample.

**Our contribution.** In this work, we propose a new seeding stage based on the sweep line paradigm [2], that allows to efficiently apply sophisticated isotopic pattern detection on large data sets. To our best knowledge, the presented algorithm is the first technique that fully exploits the two-dimensional information contained in LC-MS maps with an efficiency that scales to real-world applications. In addition, the presented method does neither rely on lossy signal pre-processing steps (baseline subtraction, noise reduction) nor on potentially disturbing resampling of the often unequally spaced data sets. Noise and baseline removal are implicitly and reversibly included through the use of a novel *isotope wavelet*, and all steps of the algorithm have been specifically designed to work even on unequally spaced spectra.

We show that our algorithm results in the selection of fewer unnecessary seeds while achieving the same accuracy as the significantly slower high-precision approach presented in [6]. On the other hand, the presented approach does not only reduce the number of false positives that have to be filtered out in the later stages of the algorithm, but also has the potential to detect true isotopic patterns that would most probably be ignored in any merely intensity-based seeding approach: as we will demonstrate, the isotope wavelet transform often identifies even hardly noticeable patterns that nearly vanish in the noise. This is particularly important at the tails of the elution profile of a given peptide where its signal intensity is low. In addition, the isotope wavelet allows for a rapid, yet very accurate classification of any isotopic pattern candidate into one of several possible charge states. Using this information, the subsequent model fitting stage of our algorithm becomes (a) significantly faster since fewer charge states have to be tested and (b) simpler and safer since a sensible initial solution for the fitting process is already provided.

The presented algorithm has been developed using OpenMS [9], a software library for shotgun proteomics, and is available to the community under the GNU Lesser General Public License.

## 2 Methods

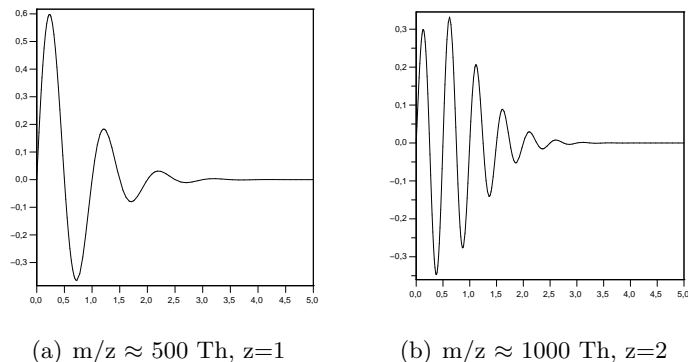
We follow the proposed sequence of seeding, region finding (extension) and refinement stage. But in contrast to previous approaches we employ a model-driven technique to detect isotopic patterns in multiple scans which will be introduced in the next two sections. Inspired by the sweep line paradigm, we collect supporting information from neighboring scans to increase our confidence in the detection. We conservatively extend this initial guess for the pattern region in the extension phase. Finally, in our refinement step, we fit a theoretical model to this region. This model is two-dimensional and consists of an average isotopic distribution of a peptide with a given mass (for the mass/charge domain) and an exponentially-modified Gaussian (for the time domain). The quality of this fit reflects our confidence in this peptide candidate and regions of low quality are discarded.

**Modeling isotope distributions of peptides.** The chemical elements of peptides naturally occur in different isotopic variants. The mass differences between light and heavy isotopes can be approximated by multiples of  $\delta_{\text{av}} \sim 1.00235$  Da [7]; for our data of intermediate resolution even  $\delta_{\text{av}} = 1$  can be used. Thus, we can compute the theoretical spectrum of a peptide from its empirical formula. There exist several algorithms for this task [10, 24, 28]. Here we use a straight-forward algorithm, the only noteworthy detail being that e.g. the isotopic distributions of  $C, C_2, C_4, C_8, \dots$  are computed by ‘squaring’. The abundances of heavy isotopes are approximated using an average amino acid, sometimes called ‘averagine’ which represents the amino acid composition observed in large protein databases. The averagine [25] has the molecular formula  $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$  and a total mass of 111.1254 Da. Fractional numbers of atoms are rounded to the next integer. We used protein sequences in a recent Swiss-Prot release to estimate the averagine composition and therefore our formula differs slightly from previous works [25].

**Wavelets to detect regions of interest.** Detecting isotopic patterns in a typical mass spectrum is complicated by the large degree of disturbing influences such as a signal baseline as well as electrical and chemical noise. Many efforts have been made to develop techniques for the automated or manually assisted reduction of these parasitics, but all of those ultimately lead to a certain distortion of the ‘real’ signal. In addition, it is often unclear if all significant noise contributions have been successfully filtered out or if some remainder has still leaked into the signal. An alternative route to identify features of interest *without explicit* removal of disturbing influences relies on signal theoretic analysis of the mass spectrometric scan, typically based on wavelets. The wavelet transform naturally generalizes the well-known Fourier transform in that a signal is *locally* split into components of different frequencies. This spectral decomposition can be very useful for feature detection algorithms, since, in reality, baseline, electrical noise, and the ‘real’ signal usually live on different frequency ranges. By computing a wavelet transformed version of the signal that corresponds to the ‘correct’ frequency range, disturbing components are automatically suppressed, greatly simplifying the analysis.

Wavelet techniques involve an additional degree of freedom in the choice of the analytical form of the so-called ‘mother wavelet’ the transform is based on. Intuitively, we replace the mass signal by a measure of how well it locally correlates with the shape of the wavelet. In [11], we have demonstrated that the established Marr wavelet is well suited for the detection of individual peaks and thus is a sensible foundation for a high-accuracy peak picking application. In particular, if the scale of the wavelet transform (corresponding to the frequency range considered) is chosen carefully, the Marr wavelet transform of a given peak is to a large degree independent of neighboring peaks, even if they strongly overlap. But while this property allows for astonishing accuracy in the separation of overlapping peaks, it is clearly counterproductive for the detection of isotopic patterns, where we explicitly want to base our analysis on the behaviour of the spectrum in the neighbourhood of a given peak. Therefore, we designed tailored ‘isotope wavelets’, which are based on the mass distributions in typical isotopic patterns as described above.

There exist other applications of the wavelet transform to mass spectrometry data [4, 23]. For example, [23] use wavelets to denoise the spectra prior to database searching for the inference of the amino acid sequence. [4] employ wavelet analysis to detect single peaks in low resolution spectra but not clearly resolved isotopic pattern for quantification as we do. However, our isotopic mother wavelet is clearly novel and has not been used before. Furthermore, we are not aware of any other work in which wavelet analysis is used for the quantification task. make Since



**Fig. 1.** Isotopic wavelets

the shape of the isotopic patterns depends on the mass as well as on the charge of the considered peptide, several wavelet functions must be used. While the wavelet can adapt automatically to the considered mass during the computation of the transform, each charge state requires its ‘own’ mother wavelet. Therefore, if we assume peptides to be at most 4-fold charged, e.g., we have to compute four transformed versions of each scan. The actual wavelet design process is technically cumbersome and out of scope of this work. Instead, we show two exemplary isotope wavelets in Fig. 1.

With the isotope wavelet, regions of interest can now be detected by first searching for local maxima in the transform. A ‘real’ isotopic pattern will lead to a chirp-like signal in the wavelet transform, since each of its mass peaks will lead to a resonance with the wavelet, and we make extensive use of this special shape to improve the specificity of our approach. Shape and regularity of the wavelet transform of a candidate pattern are used to derive a score, denoting how well the candidate fits the current wavelet type. Repeating the transform with an isotope wavelet of each charge state yields a set of charge dependent scores, leading to a powerful and robust charge prediction method.

Due to the design of the wavelet, we only need to compute one scale of the wavelet transform, i.e. compute the correlation integral of the isotope wavelet with the mass signal. While at first glance this seems to require  $\mathcal{O}(n^2)$  operations, the real runtime is actually much smaller: the wavelet has finite and typically small support that is independent of the length of the mass signal, so that the transform can be performed in linear time.

**Region extension using the sweep line paradigm.** The sweep line algorithm is a general paradigm from the field of computational geometry that has, e.g., been used to detect intersections of line segments in an efficient manner [2]. The algorithm can be illustrated by an imaginary line sliding over the segments. It keeps track of segments it encounters using a dynamic data structure. The beginning or end of a line segment triggers an update of the data-structure and a check for intersections is performed if the algorithm meets the endpoint of a segment.

We follow the general sweep line paradigm, but compared to the line segment intersection algorithm we are not searching for intersecting lines but adjacent and possibly overlapping isotopic patterns. Hence, we sweep across the LC-MS map scan by scan and use our isotope wavelet to detect the starting positions of isotopic patterns in each spectrum. That is, we apply the transform to each scan and sweep across the time domain. A significant signal in the wavelet

transform triggers an event and we check if we detected a pattern in the previous scan at the same mass with a small tolerance. The predicted monoisotopic masses of each pattern in each scan are stored in a tree-based data structure.

This approach allows to quickly discard potential isotopic peaks that are not supported by isotopic peaks in adjacent scans and to significantly reduce the number of candidate regions for the next refinement step. Furthermore, our wavelet function gives us an initial guess for the charge state of the peptide, which further reduces the number of peptide models that need to be tested. Since we scan through the LC-MS map in a linear manner, we can work efficiently on secondary memory data structures storing the peak data, allowing to apply our algorithm to very large data sets.

**Refinement stage by model fitting.** As stated above, we fit a two-dimensional model to each potential peptide signal identified in the two previous steps of the algorithm. The part of the model which is applied to the  $\frac{m}{z}$  domain relies on the average isotopic distribution for the given mass region. In addition, we model the imprecision of the mass analyzer by a normal distribution with variance  $\sigma^2$ . The resulting model for the isotopic distribution of a peptide is

$$\phi(m) = \frac{A}{\sqrt{2\pi\sigma^2}} \sum_{i=0}^{i_{\max}} a_i(m_0) e^{-(m-m_0-i\delta_{av})^2/(2\sigma^2)},$$

where  $m_0$  = monoisotopic mass,  $a_i(m_0)$  = relative abundance of  $i$ -th isotopic peak of a peptide with monoisotopic mass  $m_0$ ,  $i_{\max}$  = last isotopic peak considered, and  $A$  = area under curve.

The elution profile of the peptide is modeled by an exponentially modified Gaussian (EMG). For computational efficiency, we use a simplified version [3]. Its density function is given by

$$\text{emg}(x) = \frac{hw}{s} \sqrt{2\pi} \frac{\exp\left(\frac{w^2}{2s^2} - \frac{x-z}{s}\right)}{1 + \exp\left(-\frac{2.4055}{\sqrt{2}} \left[\frac{x-z}{w} - \frac{w}{s}\right]\right)}$$

where parameter  $z$  controls the centroid,  $w$  the width, and  $h$  the height of the elution profile. The parameter  $s$  represents the skewness of the distribution. An earlier version of our algorithm modeled the elution profiles by a normal distribution. Using the EMG makes our model much more robust in the presence of heading or tailing effects. Since the partial derivatives of the EMG can be computed analytically, we use the Levenberg-Marquardt algorithm [13, 19] to minimize the least squared distance of the model to the selected region of the LC-MS map. The quality of the model fit is measured using the squared correlation between data and model. If the correlation is too low, we discard the corresponding peptide region. The monoisotopic mass is estimated from the theoretical isotope distribution and the coordinate in retention time is taken as the centroid of the fitted EMG.

**Greedy Separation of Overlapping Isotopic Patterns.** In high-resolution spectra of complex samples, overlapping isotopic patterns might pose a severe problem for an accurate quantification. Here, we propose a greedy approach to this problem. If the wavelet transform detects overlapping isotopic patterns, these are independently assembled during the swepline stage and passed to the model fit. Peaks that have a good correlation with the theoretical model are removed from the spectrum. If the wavelet indicates that there might be another isotopic pattern in the same region, a new model is fitted using the remaining peaks in this area until no more isotopic pattern remain.

This approach is straightforward and similar to other methods already published e.g. in [7]. More sophisticated approaches are imaginable and easy to integrate into our framework. However, we found that his greedy approach works well in practice.

### 3 Results

We evaluate our approach on two different data sets. The first one was obtained from a peptide standard mix consisting of nine peptides. Here, we systematically introduced noise into the data set to evaluate the ability of our algorithm to correctly detect and quantify peptides in noisy signals.

The second data set consists of human blood serum samples from a myoglobin quantification study [20]. We use this data to show that we are able to perform very precise quantification of peptides in complex samples. Furthermore, we demonstrate that we are considerably faster than an approach which is of high accuracy but slow since it selects the seeding regions in the LC-MS map based on their intensity only [6] and therefore performs many time-consuming refinement steps using a theoretical peptide model.

**Stability analysis.** As a first step, we show that our model-driven quantification approach (*SweepWavelet*) is able to produce reliable results in the presence of noise. We systematically introduce noise in an LC-MS map of standard peptides. We are aware that performance evaluation on simulated datasets has its caveats. However, by doing so we can measure performance on data with specific characteristics.

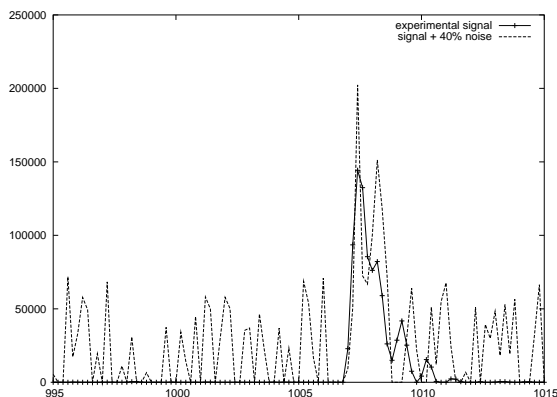
The data set chosen – an artificial mix of 9 peptides that is described in detail in [11] – is of very high quality with unusually low noise level. We thus consider manual annotation of the unperturbed spectrum as the gold standard against which we test our technique. In order to relate the noise level to the intensity of the isotopic pattern of interest, we add uniformly distributed noise with zero mean and an amplitude of 10%, 25%, 50%, and 75% of the intensity of the monoisotopic peak, respectively. If this resulted in negative values, these were replaced by zero. While this uniformly distributed noise does not provide a realistic model of all noise effects in a real spectrum, this experiment should still give us with an idea of the stability of our method if applied to noisy data.

The combination of the isotopic wavelet for seeding with the sweep line approach leads to a robust feature detection (see Table 1). Even for extremely low signal-to-noise ratios, the pattern is usually detected in a sufficient number of scans to allow for accurate seeding, and correct charge prediction. The results are exemplarily discussed for two of the peptides, with charges of 1 and 2, respectively. Performance on the other peptides is very similar. An example where the drastic amount of noise leads to severe distortion of the signal without hurting our seeding and charge prediction can be found in Fig. 2.

It is of course not clear whether our approach to introduce noise into the data comes close to real noise in mass spectra. Other possibilities would be to introduce additional isotope distributions to simulate chemical noise or to model instrument noise by adding single, poisson-distributed peaks. We decided to distort existing isotope peaks since we wanted to test the ability of our algorithm to detect deformed isotopic patterns. Adding further, artificial patterns would merely increase the running time of our algorithm and not give any information about its performance in the presence of noise. Single peaks caused by instrument noise would be simply filtered out during the wavelet transform as long as they don't resemble isotopic pattern by chance.

	Oxytocine, 1007.5 Th, charge 1					Substance P, 674.5 Th, charge 2				
	0%	10%	25%	50%	75%	0%	10%	25%	50%	75%
#scans	11/11	11/11	10/11	10/11	0/11	16/20	16/20	13/20	12/20	13/20
charge	✓	✓	✓	✓	n/a	✓	✓	✓	✓	-

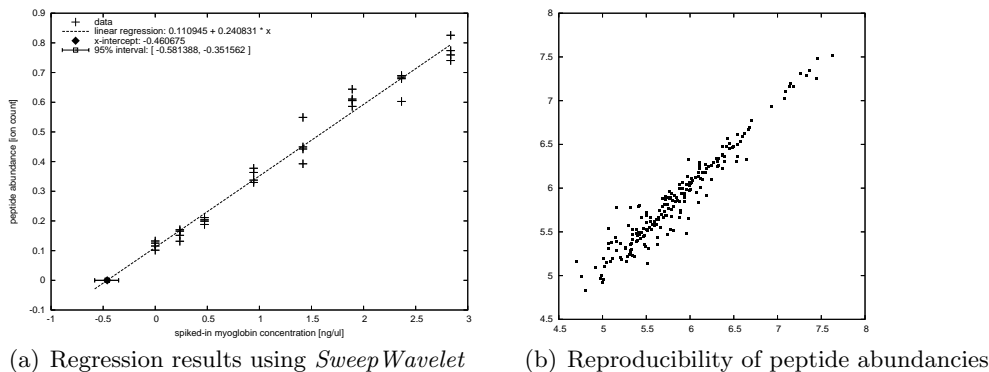
**Table 1.** Feature detection on a spectrum with varying levels of uniform noise. The percentages denote the amplitude of the noise in terms of the intensity of the monoisotopic peak of the pattern. The '#scans' - row gives the number of scans (retention times) in which the isotopic pattern was found as compared to the number found by manual annotation in the unperturbed spectrum. The 'charge' - row indicates whether the charge state was correctly assigned.



**Fig. 2.** Experimental signal with artificial noise

**Accuracy and speed of quantification.** We now apply our algorithm to a realistic task, the quantification of myoglobin from human blood serum. Myoglobin is a protein of low-molecular weight which appears quickly in blood after tissue injuries and is considered as an important biomarker for myocardial necrosis. A fast but accurate quantification of myoglobin in human blood samples is therefore important. Sample preparation and details of the absolute quantification process have already been described elsewhere [20]. In short, the myoglobin was separated from the highly abundant serum proteins by anion-exchange chromatography. The myoglobin fraction was trypsinized and the resulting peptides were analyzed by reversed-phase liquid chromatography coupled to an ion-trap mass spectrometer. To perform an absolute quantification, known amounts of human myoglobin were added to aliquots of the sample. Absolute quantification was then performed by determining the x-intercept of a linear regression using the ratio of the eleventh tryptic myoglobin peptide and an internal standard consisting of the tenth tryptic peptide of horse myoglobin. The LC-MS maps were recorded using a quadrupole ion trap mass spectrometer (Bruker Daltonics, Germany) coupled to a reversed-phase HPLC column.

We compare our approach to an algorithm [6] which was developed for this myoglobin quantification study. This method (from now on referred to as *CompLife05*) carefully collects regions of data points with high ion count and fits a theoretical isotopic model to the data. This refinement step is similar to ours but uses a simpler model. Regions having a sufficiently good correlation with the theoretical model are considered as true peptides and quantification is performed by summing the ion counts of all raw data points in the chosen region. Outlier points are excluded before quantification if their predicted intensity under the model is below a given threshold. We consider this algorithm as representative for the common approach that detects



**Fig. 3. Accuracy and reproducibility of quantification** Figure (a) shows the additive measurement as it was computed using our sweep line based algorithm. The regression was performed using the ratio of the eleventh tryptic myoglobin peptide and an internal standard consisting of the tenth tryptic peptide of horse myoglobin. (b) gives the log-transformed peptide abundances estimated from two replicate myoglobin samples.

potential peptides based on the intensity of single, but high, peaks in the LC-MS map, collects a cluster of raw data points and then refines this selection by fitting a theoretical peptide model to the data.

	SweepWavelet	CompLife05	Manual
Myoglobin data set 1	<b>True concentration</b> [ng/ $\mu$ l] <i>0.463</i>		
Computed concentration [ng/ $\mu$ l]	0.460	0.474	0.382
95% confidence interval [ng/ $\mu$ l]	[0.351;0.581]	[0.408;0.545]	[0.315;0.454]
Relative deviation from true value [%]	-0.65	+2.46	-17.42
Myoglobin data set 2	<b>True concentration</b> [ng/ $\mu$ l] <i>0.456</i>		
Computed concentration [ng/ $\mu$ l]	0.432	0.502	0.420
95% confidence interval [ng/ $\mu$ l]	[0.309;0.572]	[0.381;0.640]	[0.305;0.535]
Relative deviation from true value [%]	-5.55	+10.10	-7.89

**Table 2.** Results of absolute myoglobin quantification in human plasma. *SweepWavelet* refers to our algorithm, *CompLife05* is the approach with intensity-based seeding [6]. Column *Manual* gives the results obtained by a human expert.

Table 2 compares our algorithm to *CompLife05* and to a manual quantification by a human expert. The manual quantification was performed using the Bruker Data Analysis software and Microsoft Excel. Peak areas were estimated from extracted ion chromatograms smoothed by a Gaussian filter. The measurements were performed on two independently acquired data sets. For the computational quantification, we used the OpenMS tools [9] to perform an alignment of the LC-MS maps and to match corresponding peptides across the LC-MS maps. No smoothing or other preprocessing steps were performed.

Manual and both automated measurements estimated the true concentration of myoglobin with high precision. The regression results of *SweepWavelet* are given in Fig. 3(a). Fig. 3(b) shows a good reproducibility of the peptide abundances in the replicate measurements of the myoglobin study. Note that we do not claim to perform a significantly better quantification than [6]. Marginal differences like the ones presented above might be caused by favorable parameter settings. But we do claim that we are able to perform a quantification of equal accuracy at a much higher speed compared to a high-accuracy algorithm that was developed and tailored

for the myoglobin quantification task. Our approach is therefore more suitable for large-scale studies and high-throughput experiments.

The run time of our algorithm on a set of myoglobin maps was measured on a 3.2 GHz Intel Xeon CPU with 3 GB memory running Debian Linux (Table 3). We used the same parameter settings as for the myoglobin quantification described above. Peptide models up to charge 4 were fitted, i.e. we discarded the charge prediction of the wavelet transform to obtain a fairer comparison. Algorithm *SweepWavelet* discarded all isotopic pattern that occurred in less than three consecutive scans and *CompLife05* considered all signals up to an ion count of 4000 as potential seeds. Thus very weak peptide signals were discarded. Each data set consisted of about 1830 scans measured in full scan mode ( $\frac{m}{z}$  range 500 - 1500 Th).

The new sweep line algorithm is faster on all data sets while *CompLife05* detects about 6 times more seeds. Since the subsequent refinement step takes considerable time, *CompLife05* is significantly slower. Nevertheless this refinement step helps to discard a large number of seeds in both algorithms. Note that the number of peptides found by *CompLife05* is always higher than the number of peptides found by *SweepWavelet*. This has two reasons: the refinement step in both algorithms is imperfect and a high number of seeds will necessarily result in a higher number of false positives. Manual inspection of the results confirmed this. Second, our wavelet apparently fails to detect poorly resolved regions that show no isotopic pattern. But since mass spectrometers are evolving rapidly, we anticipate that high-resolution instruments will become standard very soon and this disadvantage will diminish.

Note that the seeds in Table 3 correspond to putative peptide signals identified either by a combination of isotopic wavelet and sweepline algorithm or based on their ion count by algorithm *CompLife05*. Column *peptides* gives the number of signals that were classified as peptide charge variants for each algorithm. *SweepWavelet* detects 200 peptides on average whereas *CompLife05* finds 500. A theoretical digest of Human Myoglobin yields only 19 peptides. The fact that both algorithms claim to find a much larger number of peptides than one would expect can be explained by several facts. The Myoglobin was extracted from human plasma. Some other peptides or contaminants will inevitably remain in the sample even after depletion and filtering. Some peptides occur in different charge states and will be independently reported by each algorithm. Finally, some signals will be false positives.

In this particular application, the high number of putative peptides was not a problem since we performed the quantification using only two Myoglobin peptides of known mass. We align the maps and filtered for the masses and expected retention times of these two peptides. In more complex applications, such as a difference detection in complex samples [16], normalisation and statistical testing for differential expression are likely to eliminate these false positive signals.

Increasing the correlation threshold in the least-squares fitting stage of both algorithms might decrease the number of false positives but also the probability of missing important signals in large-scale applications. Note that this would not influence the running time since it is mainly determined by the number of seeds on which the model fitting is performed.

Data set	SweepWavelet			CompLife05		
	Time [min]	# Seeds	# Peptides	Time [min]	# Seeds	# Peptides
Myoglobin 01	4.11	511	261	21.15	2652	521
Myoglobin 02	4.35	561	301	25.26	3537	557
Myoglobin 03	4.17	538	297	20.41	2549	492

**Table 3.** Running time, number of seeds and number of peptides after refinement on three exemplary data sets of the myoglobin study.

## 4 Conclusions and outlook

In this work, we have presented a novel algorithm for the peptide quantification problem. It combines the sweep line paradigm and a tailored wavelet function to scan for isotopic patterns in mass spectra. We have shown that this approach is able to accurately detect monoisotopic masses and charge states of peptides even in the presence of noise and that we can perform quantifications in complex data sets with high accuracy (less than 0.65% and 5.55% relative error) in an efficient manner.

Basing the feature detection process on an integral transform has a number of important advantages from a signal theoretic point of view, but might also be seen as a possible shortcoming of our algorithm: if the resolution of the data falls below a certain critical threshold, the approach is no longer practical. In our experience, however, our technique works well on real-world data. In addition, the resolution of available mass spectrometric data is ultimately going to increase, while for poorly resolved data, slower techniques like the one presented in [6] can typically be applied since the resulting maps are considerably smaller than the ones considered here.

We focused on a label-free setting in which detected peptides have to be mapped across data sets using additional computational tools. However, our approach is flexible and can also be applied to experiments in which isotope or mass tag labeling of peptides is applied. In this scenario, we have to search for pairs of peptides in the same map – a computational problem which can be solved by a range query for each detected peptide.

The application of our method presented here is the quantification of peptides from LC-MS samples. But one can easily imagine other likewise important applications such as the accurate generation of mass and time tags and peptide mass mapping.

To summarize, instead of a tiresome and error-prone manual inspection, efficient algorithms as the one presented here allow high-throughput studies and will emerge as a useful computational tool in quantitative proteomics.

## 5 Acknowledgements

We would like to thank Prof. Christian Huber (Saarland University, Saarbrücken) for providing the peptide standard mix data and the myoglobin data sets. Bettina Mayr (Saarland University, Saarbrücken) performed the manual myoglobin quantification. We are also indebted to Jens Joachim and Marcel Grunert who helped to implement an earlier version of the peptide quantification algorithm. Ole Schulz-Trieglaff was supported by the Max Planck Research School for Computational Biology and Scientific Computing Berlin. Rene Hussong and Andreas Hildebrandt were supported by DFG grants BIZ4:1-4. Clemens Gröpl and Knut Reinert acknowledge funding by the Berlin Center for Genome Based Bioinformatics (BCB). We also thank the anonymous reviewers who helped to improve this document.

## References

1. M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C.-W. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. McIntosh. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, 22(15):1902–1909, 2006.
2. J. L. Bentley and T. Ottmann. Algorithms for reporting and counting geometric intersections. *IEEE Trans. Comput*, C28:643–647, 1979.
3. Valerio B. Di Marco and G. Giorgio Bombi. Mathematical functions for the representation of chromatographic peaks. *Journal of Chromatography A*, 931:1–30, 2001.

4. Pan Du, Warren A. Kibbe, and Simon M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22:2059 – 2065, 2006.
5. B. Fischer, J. Grossmann, V. Roth, and W. Gruissem. Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics*, 22:e132–e140, 2006.
6. C. Gröpl, E. Lange, K. Reinert, O. Kohlbacher, M. Sturm, C. G. Huber, B. Mayr, and C. Klein. Algorithms for the automated absolute quantification of diagnostic markers in complex proteomics samples. In Michael Berthold, editor, *Proceedings of CompLife 2005*, Lecture Notes in Bioinformatics, pages 151–163. Springer, Heidelberg, 2005.
7. David M. Horn, Roman A. Zubarev, and Fred W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11(4):320–332, April 2000. Seminal paper on quantification of peptides of high-resolution spectra.
8. A. Keller, A. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, 74:5383–5392, 2002.
9. O. Kohlbacher, K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm. TOPP - The OpenMS proteomics pipeline. In *Proceedings of the 5th European Conference on Computational Biology (accepted)*, 2006.
10. H. Kubinyi. Calculation of isotope distributions in mass spectrometry. a trivial solution for a non-trivial problem. *Anal. Chim. Acta*, 247:107–119, 1991.
11. E. Lange, C. Gröpl, K. Reinert, O. Kohlbacher, and A. Hildebrandt. High accuracy peak-picking of proteomics data using wavelet techniques. In *Proceedings of the Pacific Symposium on Biocomputing (PSB) 2006*, pages 243–254, 2006.
12. K. C. Leptos, D. A. Sarracino, J. D. Jaffe, B. Krastins, and G. M. Church. MapQuant: Open-Source software for large-scale protein quantification. *Proteomics*, 6(6):1770–1782, 2006.
13. K. Levenberg. A method for the solution of certain problems in least squares. *Quart. Appl. Math.*, 2:164–168, 1944.
14. H. Li, X.-J. and Zhang, J.R. Ranish, and R. Aebersold. Automated statistical analysis of protein abundance ratios from data generated by stable isotope dilution and tandem mass spectrometry. *Anal. Chem.*, 75:6648–6657, 2003.
15. X.-J. Li, E. C. Yi, C. J. Kemp, H. Zhang, and R. Aebersold. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell Proteomics*, 4(9):1328–1340, 2005.
16. J. Listgarten and A. Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics*, 4(4):419–434, 2005.
17. M.J. MacCoss and D. E. Matthews. Quantitative MS for proteomics: Teaching a new dog old tricks. *Anal. Chem.*, 77(15):294A–302A., 2005.
18. M. Mann and R. Aebersold. Mass spectrometry-based proteomics. *Nature* 422, 422:198 – 207, 2003.
19. D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441, 1963.
20. B. M. Mayr, O. Kohlbacher, K. Reinert, M. Sturm, C. Gröpl, E. Lange, C. Klein, and C. Huber. Absolute myoglobin quantitation in serum by combining two-dimensional liquid chromatography-electrospray ionization mass spectrometry and novel data analysis algorithms. *J. Proteome Res.*, 5:414–421, 2006.
21. S.-E. Ong and M. Mann. Mass spectrometry-based proteomics turns quantitative. *Nature Chem. Biology*, 1(5):252–262, 2005.
22. P. M. Palagi, P. Hernandez, D. Walther, and R. D. Appel. Proteome informatics I: Bioinformatics tools for processing experimental data. *Proteomics*, ISSN 1615-9861, <http://dx.doi.org/10.1002/pmic.200600273> (epub ahead of print), 2006.
23. Tomas Rejtar, Hsuan shen Chen, Victor Andreev, Eugene Moskovets, and Barry L. Karger. Increased identification of peptides by enhanced data processing of high-resolution maldi tof/tof mass spectra prior to database searching. *Analytical Chemistry*, 76:6017 –6028, 2004.
24. A.L. Rockwood, S.L. Van Orden, and R.D. Smith. Rapid calculation of isotope distributions. *Analytical Chemistry*, 67:2699–2704, 1995.
25. M. W. Senko, S. C. Beu, and F. W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, 1995.
26. C. A. Smith, E. J. Want, G. O’Maille, R. Abagyan, and G. Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006.

27. G. Wang, W.W. Wu, T. Pisitkun, J.D. Hoffert, M.A. Knepper, and R.-F. Shen. Automated quantification tool for high-throughput proteomics using stable isotope labeling and LC-MS. *Analytical Chemistry*, 78(16):5752–5761, 2006.
28. J.A. Yergey. A general approach to calculating isotopic distributions for mass spectrometry. *Int. J. Mass Spectrom. Ion Phys.*, 52:337–349, 1987.