Sparse Deconvolution for Peak Picking and Ion Charge Estimation in Mass Spectrometry

Kristian Bredies[1], Theodore Alexandrov[1], Jens Decker[2], Dirk A. Lorenz[1], and Herbert Thiele[2]

[1] Center for Industrial Mathematics, University of Bremen, D-28334 Bremen, Germany, {kbredies, theodore, dlorenz}@math.uni-bremen.de
[2] Bruker Daltonik GmbH, D-28359 Bremen, Germany, {jens.decker, herbert.thiele}@bdal.com

Summary. In this paper we propose a new procedure for peak detection in mass-spectrometry data using sparse deconvolution. We apply the procedure for estimation of the ion charges for isotopic patterns of overlapping peaks. The evaluation is performed on the thymosin β_4 16-38 fragment measurements. Moreover, a comparison with the Mexican hat based algorithm of peak picking is provided.

1 Introduction

For mass spectrometry (MS) the detection of m/z (i.e. mass over charge) peaks is a vital step of the data processing pipeline. The purpose of a MS peak picking algorithm is the transformation of a profile spectrum into a list of peaks. For most instruments the profile spectra are obtained by digitalization of a time dependent signal.

The main required properties of a peak detection algorithm are: (1) good m/z precision and accuracy, (2) resolution of overlapping peaks, (3) selective recognition of noisy peaks and (4) performance.

The resolution of overlapping isotopic peaks (for example see Fig. 1) is important to resolve overlapping chemical compounds and to determine the distance between the isotopic peaks of the same molecular ion. The distance between two isotopic peaks in the pattern for the charge z is approximately $1.00235/z$ Th for peptides with deviations in the milli-Thomson range depending on the exact formula. Based on this distance it is possible to determine the charge of an ion which is e.g. important for the real time selection of the most promising ions for fragmentation experiments. This is especially the case for ion trap instruments which have a typical full width half maximum of 0.2–0.5 Th depending on the measurement mode and the type of instrument. Even though the peak resolution is rather limited, these instruments are still of large interest especially due to their large sensitivity and comprehensive fragmentation capabilities.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_43,

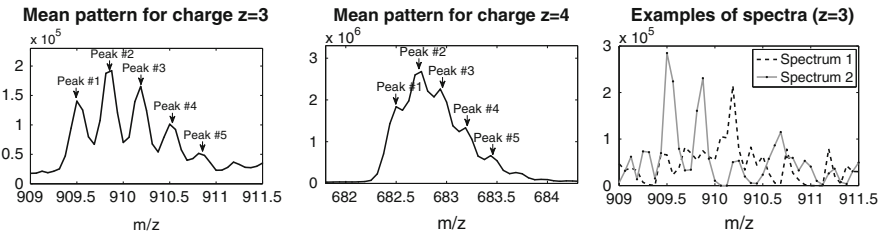

Fig. 1. Mean isotopic patterns of overlapping peaks for ion charges $z = 3$ and $z = 4$ and two examples of spectra for the ion charge $z = 3$

1.1 Mathematical Formulation of the Problem

The problem under study is (1) detection and picking of overlapping isotopic peaks and (2) estimation of the charge of the molecular ion using the distance between the isotopic peaks found. The distance between two neighboring isotope peaks can be assumed to be $1/z$ Th which allows a determination of the charge z provided that the positions of the peaks are known.

For finding the actual positions of the peaks for one isotope pattern, the model assumption is that the measured data f is composed of only few peaks of a known shape G_σ, i.e.

$$f = \sum_{i \in I} u_i G_{\sigma,i} \ .$$

We suppose $G_{\sigma,i}$ to be a Gaussian peak at position i whose area has been normalized to 1: $G_{\sigma,i}(x) = c_\sigma \exp\bigl(-\sigma(x-i)^2\bigr)$. Its width can be tuned by σ and is usually given by the characteristics of the utilized mass spectrometer and resolution. Moreover, u_i represent the corresponding coefficients and I is the (finite) collection of positions we are considering for the peaks.

This model can be interpreted as a result of convolution of several Dirac delta peaks (of heights u_i) with the Gaussian kernel that happened in the process of mass spectrometry measurements.

2 Proposed Method

2.1 Peaks Detection Through Sparse Deconvolution

For isotope patterns, it is important to suppose that the number of actual peaks, i.e. the number of non-zero coefficients u_i, is significantly less than the number of available peaks in I. Such a model assumption can be mathematically implemented by taking so-called "sparsity constraints" into account. Recovering a series of delta peaks from convolved data is known as "sparse

deconvolution" and has been studied recently [3,4]. Moreover, sparsity assumptions can serve as a regularization to reduce the sensitivity with respect to noisy data, see [6] and the references therein. Therefore, for the deconvolution of the isotope patterns, we are following a variational approach which amounts to the solution of the following minimization problem:

$$\min_u \frac{\| \sum_{i \in I} u_i G_{\sigma,i} - f \|^2}{2} + \alpha \sum_{i \in I} |u_i|$$

In the algorithm, I consists of equally sampled points covering the part of the spectrum to be deconvolved. The sampling rate $(\Delta i)^{-1}$ is chosen to be significantly higher than the sampling rate of the spectrum, usually of the factor 4. Additionally, the regularization parameter α is set to be a multiple of the area of the data, i.e. $\alpha = \tau \sum_j |f_j|$ with a $\tau > 0$. The solution of the minimization problem was done by an iterative thresholding algorithm from [2].

2.2 Charge Estimation

This part of the algorithm takes a deconvolved isotope pattern u and tries to extract its charge z by examining the distances between the peak positions. First, it extracts the N most significant peaks, i.e. an ascending sequence of i for which the u_i correspond to the N greatest values of u. In the implemented algorithm, we chose $N = 5$. Each of the positions i_k are corrected by fitting a parabola to the coordinates $(i - \Delta i, i, i + \Delta i)$ and the corresponding values. Then i_k is replaced by the position of the parabola maximum. Subsequently, all differences between the i_k are collected and weighted according to how many peaks are skipped, i.e.

$$D = \left((i_{k+1} - i_k)_k, \tfrac{1}{2}(i_{k+2} - i_k)_k, \tfrac{1}{3}(i_{k+3} - i_k)_k, \ldots, \tfrac{1}{N}(i_N - i_1) \right) .$$

For D, the mean value m as well as the variance V are computed. The charge z is then estimated by the integer closest to $1/m$.

In order to make the charge estimation more robust, the following additional step is performed. Assuming that there is one outlier in the collection of positions $(i_k)_k$, we compute the corresponding charge estimate as well as the variance for the positions where

1. i_k is left out for $k = 1, \ldots, N$,
2. i_k is replaced by $\frac{1}{2}(i_{k+1} + i_{k-1})$ for $k = 2, \ldots, N-1$.

Eventually, the charge which corresponds to the smallest variance for the above test is returned.

Algorithm

1. Given: isotope pattern f
   - Create a set of positions $I = i_{\text{start}} : \Delta i : i_{\text{end}}$
   - Compute $\alpha = \tau \sum_j |f_j|$
2. Solve the minimization problem $\min_u \ \frac{1}{2} \| \sum_{i \in I} u_i G_{\sigma,i} - f \|^2 + \alpha \sum_{i \in I} |u_i|$
3. Extract most significant peaks
   - Find the indices $i_1, \ldots, i_N$ of the N greatest u_i
   - Replace each peak i_k by the maximum of the fitting parabola
4. Create reduced peak lists $(P_p)_p$
   - P_0: original peak list
   - $P_1, \ldots, P_N$: peak list with i_p left out
   - $P_{N+1}, \ldots, P_{2N-2}$: peak list with i_{p-N+1} replaced by $\frac{1}{2}(i_{p-N} + i_{p-N+2})$
5. For each reduced peak list: extract charge z_p
   - Compute all differences $(i_k - i_l)/(k - l)$ for all $k > l$
   - Calculate mean m_p and variance V_p of differences
   - Set $z_p = \text{round}(1/m_p)$
6. Return charge z_p corresponding to the minimal V_p

3 Experiments

3.1 Data Set Description

To get spectra with known m/z values and multiple charge states a direct injection measurement of the thymosin β_4 16-38 fragment (Bachem No. H-2926) was done using 200 fmol/µl in 50% acetonitrile, 0.1% formic acid at 3 µl/min. A HCT Ultra ETD II instrument from Bruker Daltonik was used for these measurements. In total 462 spectra were accumulated in the enhanced standard mode without moving average and prefiltering.

3.2 Peak Picking Using the Mexican Hat Wavelet

For the data set given, we compared our peak picking procedure (Algorithm steps 1–3) with a procedure based on using the Mexican hat (MH) wavelet.

Traditional peak picking algorithms are looking for a zero crossing of the 1st derivative. The detection of peaks which do not give a maximum anymore is not possible in that way. The detection of overlapping peaks also becomes difficult.

One approach to overcome this problem is to use the 3rd derivative instead of the 1st [8] to be able to detect shoulder peaks. Using higher derivatives enhances the influence of noise. For Gaussian peaks and a normal distributed noise assumption it can be shown [1] that smoothing with a Gaussian kernel is the optimal filter. Combining the calculation of the 2nd derivative with a Gaussian smoothing is equivalent to convolving the data with a Mexican hat wavelet which is the 2nd derivative of a Gaussian. This approach was successfully used by [5,7] and exploited within the OpenMS framework.

3.3 Comparison Results

The evaluation of the proposed algorithm (denoted SD hereafter) and the comparison with the Mexican hat wavelet based procedure (MH) was organized as follows. For all the spectra we detected peaks positions in two m/z regions $\mathcal{I}_3 = [909, 911.5]$ Th and $\mathcal{I}_4 = [681.5, 684.5]$ Th containing the isotopic patterns for the charges $z = 3$ and $z = 4$, respectively. For each region the distance between the peaks was calculated using our Algorithm (steps 4–6) and converted to the charge value. Then, for each region ($\mathcal{I}_3$ and $\mathcal{I}_4$) the error rate (E_3 and E_4, respectively) was calculated, i.e. the ratio (in percentage) of the correctly evaluated charges to the number of spectra. The algorithms are rated according to the mean rate $E = (E_3 + E_4)/2$.

Both SD and MH have parameters. SD, besides the number of iterations (10,000 iterations are used), has the parameters σ (manages the supposed width of the peaks) and τ (the regularization parameter). The MH algorithm has the two parameters n and c. The parameter c is defining the width of the Mexican hat function in units of the data point distance and is equal to the standard deviation of the Gaussian function defining the Mexican hat function (but not directly equal to σ in SD); $2n + 1$ is the number of data points for which the Mexican hat function is defined (0 beyond that range).

A grid search has been performed where for each pair of parameters the mean error rate was calculated. The calculated mean rates are presented in

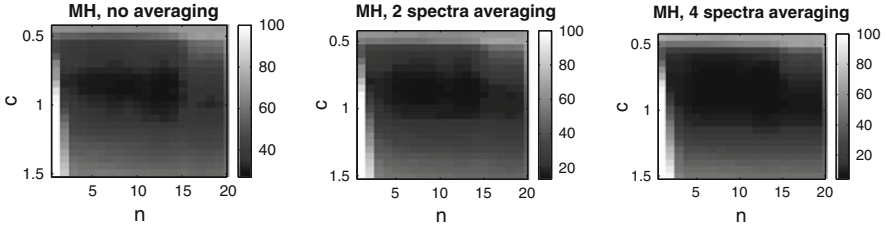

Fig. 2. Grid search results for MH: the mean error rate E (in percentage) for different pairs of parameters n and c

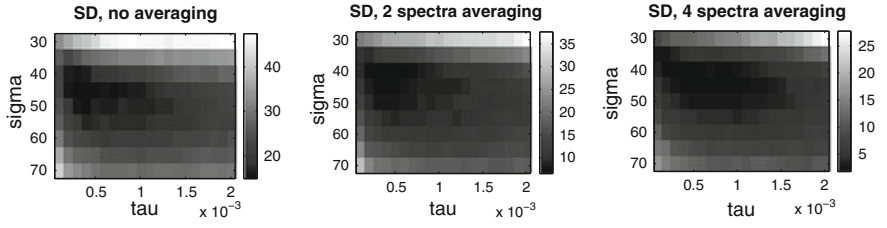

Fig. 3. Grid search results for SD: the mean error rate E (in percentage) for different pairs of parameters σ and τ

Table 1. The minimum mean error rate E (in percentage, over all tested parameters) of our algorithm (SD) and of the Mexican hat based algorithm (MH) calculated for 462 original spectra ("no averaging"), for 461 spectra resulting after averaging of each two neighbor spectra ("2 spectra averaging") and for 459 spectra after averaging of each four neighbor spectra ("4 spectra averaging")

	No averaging	2 Spectra averaging	4 Spectra averaging
SD	14.8%	6.8%	2.5%
MH	27.2%	14.3%	4.8%

Figs. 2 (MH) and 3 (SD). Table 1 (column "no averaging") contains the minimal values of the mean error rates over all tested pairs of parameters. The SD algorithm significantly outperforms the MH peak picking algorithm.

The results can be enhanced by averaging several spectra previous to the peak picking as the data is very noisy (see Fig. 1 for examples of spectra). Though the averaging requires additional measurements (technical replicates), this operation is often used in MS. We simulated the averaging by taking means of two and four neighbor spectra that reduces the data set size to 461 and 459 spectra, respectively. Table 1 contains the mean error rates computed after averaging. The averaging of four spectra for example improves the error rates by the factor of 7. Both procedures (MH and SD) provide low error rates but SD is significantly better than MH for all types of averaging used.

The only disadvantage of SD is its runtime which is 17 min versus 7 s for MH (on an Intel 2.66 GHz PC, for fixed parameters).

Acknowledgements

Thanks to Markus Lubeck, Bruker Daltonik, for doing the MS measurements.

References

1. Andreev, V., Rejtar, T., et al.: Anal. Chem. **75**, 6314–6326 (2003)
2. Bredies, K., Lorenz, D.: SIAM J. Sci. Comput. **30**, 657–683 (2008)
3. Dahlke, S., Maass, P., et al.: Mathematical Methods in Time Series Analysis and Digital Image Processing, 75–109 (2007)
4. Klann, E., Kuhn, M. et al.: Inverse Probl. **23**, 2231–2248 (2007)
5. Lange, E., Gröpl, C., et al.: Proc. Pacific Symp. on Biocomp. 243–245 (2006)
6. Lorenz, D.: J. Inverse Ill-Posed Probl. **16**, 463–478 (2008)
7. Sturm, M., Bertsch, A., et al.: BMC Bioinformatics **9**, 163 (2008)
8. Vivó-Truyols, G., Torres-Lapasió, J., et al.: J. Chromatrogr. A **1096**, 146–155 (2005)