METHODOLOGY

Open Access

CrossMark

Informed baseline subtraction of proteomic mass spectrometry data aided by a novel sliding window algorithm

Tyman E. Stanford[*], Christopher J. Bagley and Patty J. Solomon

Abstract

Background: Proteomic matrix-assisted laser desorption/ionisation (MALDI) linear time-of-flight (TOF) mass spectrometry (MS) may be used to produce protein profiles from biological samples with the aim of discovering biomarkers for disease. However, the raw protein profiles suffer from several sources of bias or systematic variation which need to be removed via pre-processing before meaningful downstream analysis of the data can be undertaken. Baseline subtraction, an early pre-processing step that removes the non-peptide signal from the spectra, is complicated by the following: (i) each spectrum has, on average, wider peaks for peptides with higher mass-to-charge ratios (m/z), and (ii) the time-consuming and error-prone trial-and-error process for optimising the baseline subtraction input arguments. With reference to the aforementioned complications, we present an automated pipeline that includes (i) a novel 'continuous' line segment algorithm that efficiently operates over data with a transformed m/z-axis to remove the relationship between peptide mass and peak width, and (ii) an input-free algorithm to estimate peak widths on the transformed m/z scale.

Results: The automated baseline subtraction method was deployed on six publicly available proteomic MS datasets using six different m/z-axis transformations. Optimality of the automated baseline subtraction pipeline was assessed quantitatively using the mean absolute scaled error (MASE) when compared to a gold-standard baseline subtracted signal. Several of the transformations investigated were able to reduce, if not entirely remove, the peak width and peak location relationship resulting in near-optimal baseline subtraction using the automated pipeline. The proposed novel 'continuous' line segment algorithm is shown to far outperform naive sliding window algorithms with regard to the computational time required. The improvement in computational time was at least four-fold on real MALDI TOF-MS data and at least an order of magnitude on many simulated datasets.

Conclusions: The advantages of the proposed pipeline include informed and data specific input arguments for baseline subtraction methods, the avoidance of time-intensive and subjective piecewise baseline subtraction, and the ability to automate baseline subtraction completely. Moreover, individual steps can be adopted as stand-alone routines.

Keywords: Mathematical morphology, Top-hat operator, Line segment algorithm, Mass spectrometry, Baseline subtraction, Pre-processing, Matrix-assisted laser desorption/ionization, Time-of-flight, Unevenly spaced data

*Correspondence: tyman.stanford@adelaide.edu.au
School of Mathematical Sciences, The University of Adelaide, North Terrace, 5005 Adelaide, Australia

Stanford *et al. Proteome Science* (2016) 14:19

Page 2 of 15

Background

Discovery of protein biomarkers by mass spectrometry

Protein biomarkers are proteins or protein fragments that serve as markers of a disease or condition biomarkers [1, 2] by virtue of their altered relative abundance in the disease state versus the healthy condition. Matrix-assisted laser desorption/ionisation (MALDI) linear time-of-flight (TOF) mass spectrometry (MS) is a widely used technology for biomarker discovery as it can create a representative profile of polypeptide expression from biological samples. These profiles are displayed as points of polypeptide abundance (intensity; the *y*-axis) for a range of mass-to-charge values (*m/z*; the *x*-axis). Each spectrum is an array of positive intensity values for discretely measured *m/z* values, but the profile is typically displayed on a continuous scale. MALDI TOF-MS spectra are typically limited to polypeptides less than 30 kilo Daltons although there is no theoretical upper limit [3]. Numerous biomarkers using MALDI TOF-MS have been identified to date [3, 4].

Statistical analysis of the proteomic profiles for biomarker discovery cannot be undertaken without prior removal of noise and systematic bias present in the raw spectra. This removal is conducted through a series of steps known as pre-processing. Pre-processing generally consists of five steps to remove false signal as set out in Fig. 1: signal smoothing, baseline subtraction, normalisation, peak detection and peak alignment. Signal smoothing and baseline subtraction are adjustments made to each spectrum individually (i.e., intra-spectrum pre-processing), while normalisation and peak alignment (after peak detection) are adjustments made to make each spectrum within an experiment comparable (i.e., inter-spectrum pre-processing).

Signal smoothing is the first step in pre-processing the data and aims to remove instrument-derived noise in the data and stochastic variation in the spectrum signal. Baseline subtraction then follows, which is the removal of the estimated 'bed' on which the spectral profile sits, composed of non-biological signal, e.g. chemical noise from ionised matrix. Normalisation is the third step in pre-processing. This has the aim of making the observed signals proportionate over the experiment; to correct for instrument variability and sample-ionisation efficiency that will influence the number of peptide ions reaching the detector. Peak detection is the fourth step, which is the detection of peak signal as peptide mass and intensity pairs. Finally, in the fifth step, the peaks are subject to peak alignment which adjusts for small drifts in *m/z* location which result from the calibration required for the TOF-MS system. This ensures that peptides common across spectra are recognised and compared at the same *m/z* value. Once the data have been pre-processed, analysis to detect potential biomarkers can be performed.

There are numerous freely available MS pre-processing packages. For example, in the `R` statistical software environment, `MALDIquant`, `PROcess` and `XCMS` are available [5–9]. Although we have set out the usual sequence of five data pre-processing steps, an optimal approach to pre-processing is not yet established and there is scope to improve current pre-processing methods and the order in which they are applied, to allow more reliable biomarker identification [10]. The present paper focuses on optimising methods for the baseline subtraction step of pre-processing of the raw spectra.
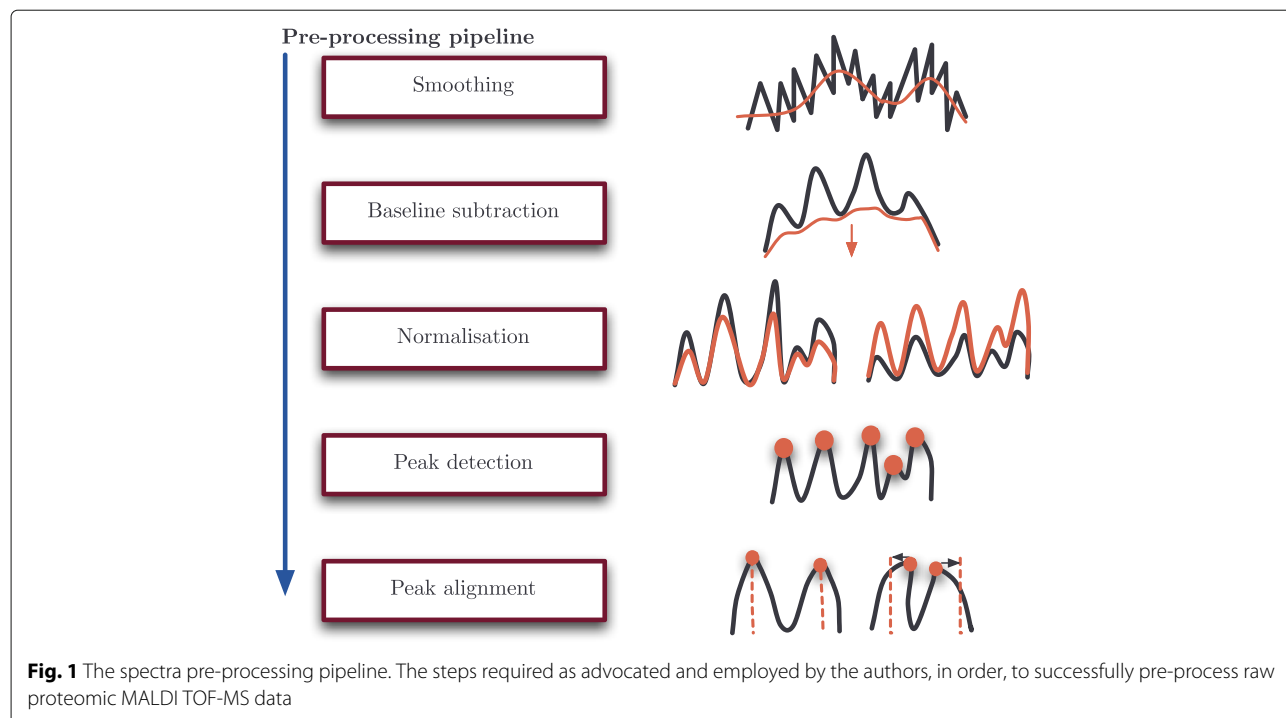
Baseline subtraction

The non-biological signal to be removed by baseline subtraction is often described as 'chemical noise' which predominantly occurs at low mass values and may result from ionised matrix molecules [11]. An example of a MALDI TOF-MS spectrum, a baseline estimate and the resulting baseline subtracted spectrum are shown in Fig. 2. The spectrum in Fig. 2 is a from the Fiedler dataset which is outlined in the 'Data used' section in Methods. The pre-processing applied prior to the baseline subtraction involved taking the square root of the spectrum intensities (for variance stabilisation) and performing the first pre-processing step in smoothing using the Savitzky-Golay method with a half window size of 50 [12].

The baseline subtraction method discussed in the present paper utilises the *top-hat* operator, which is an operator defined in mathematical morphology. Mathematical morphology was originally proposed for two-dimensional image analysis then further developed for image processing of microarray data images [13, 14]. It has since been applied to MS data [7, 15–19], and we describe the theory that is largely ignored when applied naively. The mathematical morphology definitions of an erosion, dilation, opening and top-hat allow us to extend the current use of mathematical morphology in MS baseline subtraction.
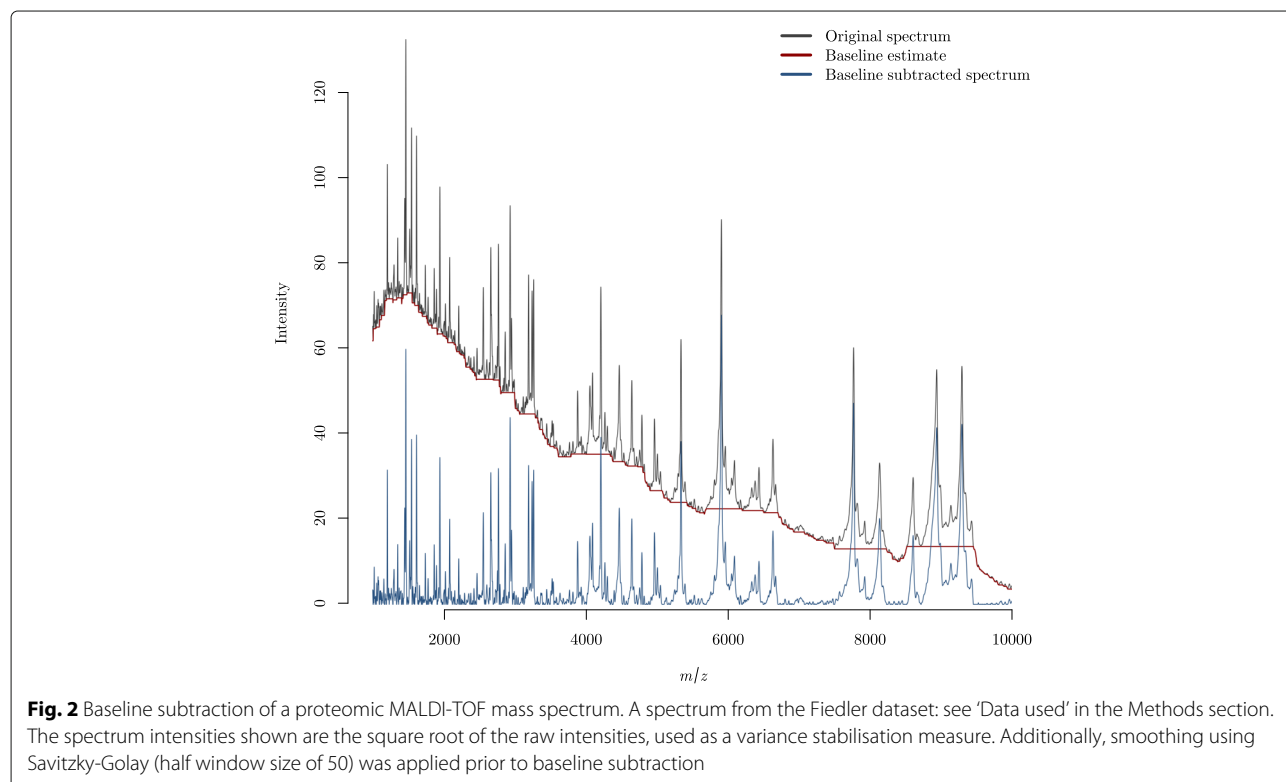
The top-hat operator has some properties, i.e. it is a non-parametric and non-linear filter, which make it desirable for baseline subtraction. In particular, this suits the non-biological signal in MS spectra which may not follow a known functional form. Furthermore, the top-hat operator is computationally inexpensive compared with standard functional filters that require estimates of model parameters.

Other algorithmic methods of baseline subtraction such as the sensitive nonlinear iterative peak (SNIP) algorithm [20, 21] provide an alternative to the top-hat operator. However, it will be shown in the Methods section that the top-hat operator can importantly be extended, using the mathematical theory underpinning it, for unevenly spaced data.

Stanford *et al. Proteome Science* (2016) 14:19

Page 3 of 15

Fig. 1 The spectra pre-processing pipeline. The steps required as advocated and employed by the authors, in order, to successfully pre-process raw proteomic MALDI TOF-MS data

Standard methods of baseline subtraction estimate local minima (troughs) and fit either local regression (LOESS, Savitzky-Golay) or interpolate (splines) through these points [22]. These methods require careful selection of the window size for detecting troughs, the polynomial order and the span of points for fitting the model, where applicable. Despite using optimised input arguments for these methods, they cannot guarantee a non-negative resultant signal without applying contraints to guarantee non-negativity. Without such constraints, padded or

Fig. 2 Baseline subtraction of a proteomic MALDI-TOF mass spectrum. A spectrum from the Fiedler dataset: see 'Data used' in the Methods section. The spectrum intensities shown are the square root of the raw intensities, used as a variance stabilisation measure. Additionally, smoothing using Savitzky-Golay (half window size of 50) was applied prior to baseline subtraction

Stanford *et al. Proteome Science*   (2016) 14:19

Page 4 of 15

removed signal in places of high curvature in the spectra may be produced. This can easily be envisioned by considering two local minimums and an adjacent point to one of the local minimums that lies between both. There is no property that stops the adjacent point lying below an interpolation of the two minimums, especially where there exists a large difference between the values of the local minimums.

The top-hat operator

The top-hat operator is a function described within mathematical morphology theory, an area that is heavily applied in image processing and analysis [23, 24]. For the interested reader, we direct you to 'Additional file 1' for a mathematical description.

The top-hat operator is the end result of applying a rolling (moving) minimum calculation, called an *erosion*, to a spectrum; then applying a rolling maximum calculation, called a *dilation*, to the erosion; and finally removing this estimate of the spectrum's baseline, called an *opening*, from the initial spectrum.

The rolling minimum and maximum calculations require an appropriately sized line-segment, or window, to be defined which provides the local domain for the minimum and maximum calculations to be made. In mathematical morphology parlance this window is called a *structuring element* (SE).

The opening has the desirable property that it is restricted to values equal or less than the spectral values to which it was applied. In turn, the top-hat operator therefore provides a background estimate and removal without risk of creating negative signal, since it is a physical impossibility of the system.

Current application of the top-hat operator to linear TOF-MS

A naive algorithmic application of an erosion to spectral intensities simply requires a traversal of each point, where the minimum value within a window over that point is the resulting erosion. The process is performed similarly for a dilation. However, erosions and dilations can be calculated more efficiently with the line segment algorithm (LSA) [25, 26]. Application of the LSA is mainly seen in medical imaging and analysis [27, 28]. The R package `MALDIquant` and `OpenMS` use this algorithm in their implementation of the top-hat operator.

When applying the top-hat operator to a spectrum, the SE needs to be chosen carefully. In particular, the following need to be considered.

1. If a SE is too large, then it will be too conservative and leave false signal.
2. If a SE is too small, it will result in under-cut peaks and remove valid signal.

3. The mean peak width increases further along the m/z-axis [29]. The baseline subtraction needs to be performed in a piecewise manner, otherwise the above issues 1 and 2 will occur.

Despite the simplicity of the top-hat operator compared to functional alternatives, piecewise baseline subtraction is still required. In fact, piecewise baseline subtraction should be applied for any method that implicitly assumes peak width remains constant, such as local regression, interpolating splines or the SNIP algorithm.

The SE size used for the top-hat operator needs to be of equivalent window size to each spectrum's peak widths, or greater, to ensure the top-hat operator does not 'under-cut' peak intensities. The piecewise baseline subtraction involves determining subsections of the m/z-axis, where fixed SE widths (in the number of m/z points) in each section are appropriate, or the equivalent input arguments for other baseline methods. Smaller SEs will be chosen corresponding to lower m/z values and larger SEs will be used corresponding to larger m/z values.

Figure 3(a) illustrates a spectrum from the Fiedler data separated into four roughly equal segments based on the number of intensity values. When applying the top-hat operator, the SE size is a constant number of intensity values within each piecewise section of the axis. The SE sizes selected in Fig. 3(a) were made by visual inspection and trial-and-error. Figure 3(b) depicts the same spectrum as Fig. 3(a) but the x-axis is in terms of m/z location. On this m/z-axis, the SE size increases along the m/z-axis within each piecewise segment simply by virte of the distances between m/z points increasing, even though the same window size is being used in terms of the number of intensity values. However, the increasing coverage in m/z units across the m/z-axis is not proportional to the increase in peak widths. Figure 3(a) and (b) demonstrate that there is not a constant number of intensity values for the SE across the entire m/z-axis that could avoid conservative baseline estimates (1) or under-cut peaks (2) or even both.

Improving baseline subtraction

Prior to pre-processing MALDI TOF-MS data, a log or square root transformation of the intensity axis is usually performed as a variance stabilisation measure but no such transformation is made to the m/z-axis. If an appropriate m/z transformation could be made however, piecewise pre-processing of the spectra for the baseline subtraction step (and potentially for other pre-processing steps) could be avoided. Additionally, the default arguments such as window size in software to perform baseline subtraction are statically defined. Uninformed default arguments such as these are highly likely to need modification for successful baseline subtraction, as spectra attributes vary from one experiment to another. Dynamic default arguments
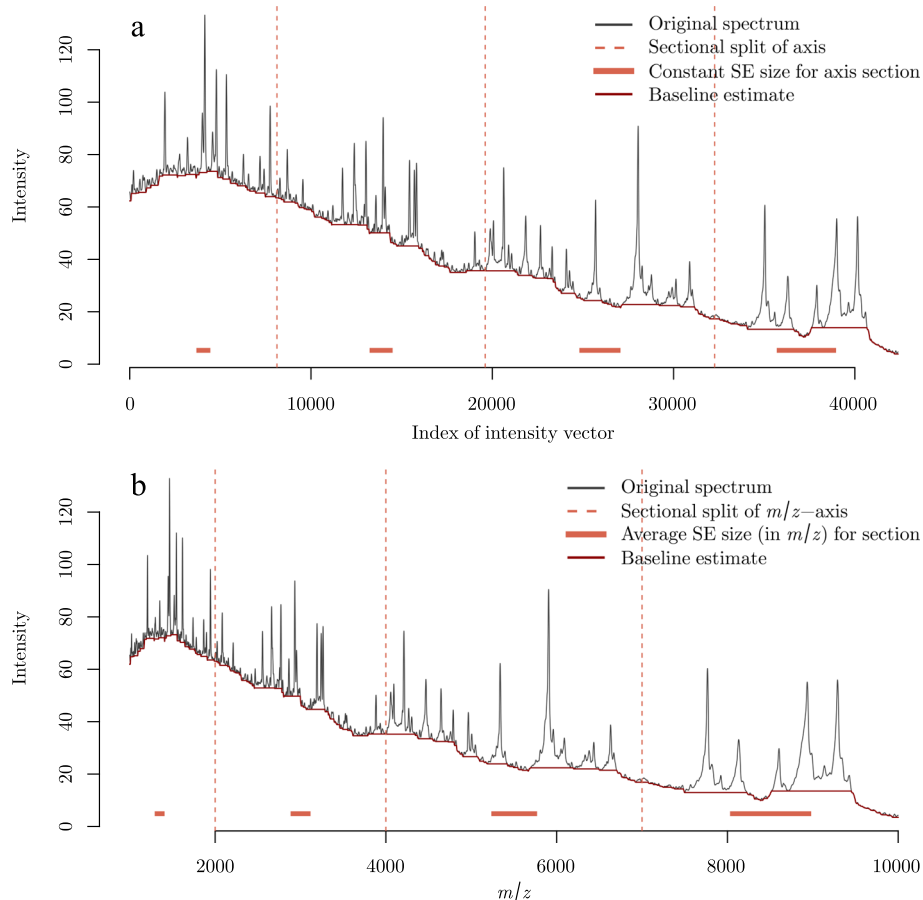
Stanford *et al. Proteome Science* (2016) 14:19

Page 5 of 15

Fig. 3 Piecewise baseline subtraction of a proteomic MALDI-TOF mass spectrum from the Fiedler dataset using the top-hat operator. An example of piecewise top-hat baseline estimation where the *x*-axis is shown as **a** the sequential number of the *m/z* point of the 42,388 *m/z* points, and **b** the *m/z* location. The window size of the top-hat operator is provided as a number of *m/z* points. Since the spacing of *m/z* values increases along the *x*-axis in (**b**), the SE size is not constant within sub-intervals of the *m/z*-axis

that are informed by the data would be an advantage in saving both user time and minimising user error.
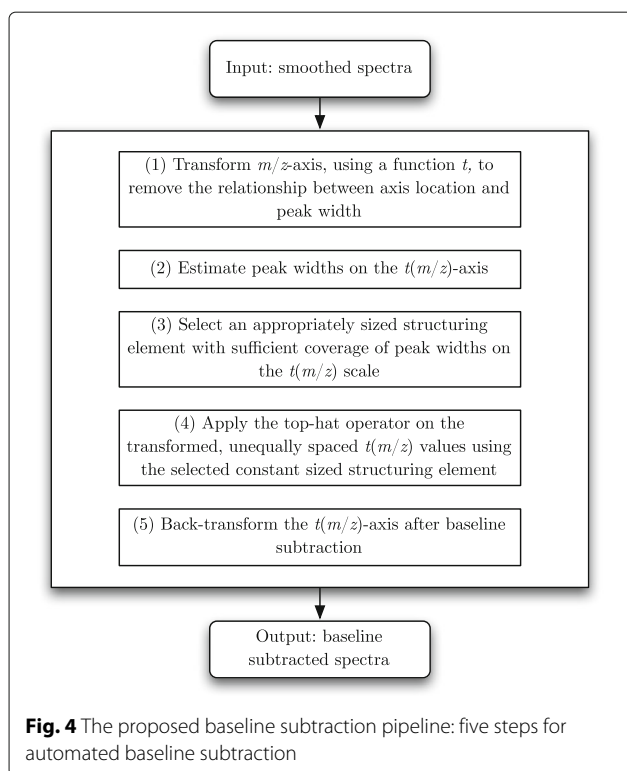
Methods

A pipeline to achieve automated baseline subtraction

The pipeline shown in Fig. 4 can be employed to automate the baseline subtraction step. The first step of the pipeline requires a suitable transformation of the *m/z*-axis. If such a transformation of the *m/z*-axis can be made, a piecewise approach is not required as a constant-sized SE can be used over the entire spectrum. A log-type transform that expands the low *m/z* values and contracts the high *m/z* values is required. Once a suitable transformation is found, a top-hat operator defined over non-evenly spaced real values (as opposed to integer values) can be used at the baseline subtraction step. The implementation requires a minimum and a maximum sliding window algorithm for unevenly spaced data which means the LSA cannot be used. Naive algorithms are available; however,

here we present a novel sliding window algorithm that we show outperforms naive sliding window algorithms by avoiding repeated minimum (or maximum) calculations for common points in successive sliding windows. However, a SE size does need to be selected. This can be implemented by firstly estimating peak widths, then selecting a SE size that covers a sufficient proportion of the estimated peak widths. The process of estimating peak widths can be automated without user input and our recommend approach is presented here. The final step in the baseline subtraction pipeline is simply the (reverse) transformation back to the original *m/z* scale.

The new pipeline to perform baseline subtraction of MALDI TOF-MS data presented in Fig. 4 has two major advantages when compared to standard methods.

- Firstly, the pipeline automates the baseline subtraction step, that is otherwise conducted in a piecewise manner. This eliminates the need for user

Stanford *et al. Proteome Science* (2016) 14:19

Page 6 of 15

Fig. 4 The proposed baseline subtraction pipeline: five steps for automated baseline subtraction

input and time-consuming calibration by observation. Automation of the baseline subtraction step also minimises the potential for user error and the time required to assess the input arguments for optimality.

- Secondly, the novel algorithm that is computationally less expensive than a naive minimum or maximum sliding window algorithm to perform the top-hat operation on unevenly spaced data, presented here, further minimises the computational time burden of baseline subtraction.

Fields of application outside of bioinformatics that encounter unevenly spaced data are also likely to find this algorithm useful in practice. Other names for unevenly spaced data include unevenly sampled, non-equispaced, non-uniform, inhomogeneous, irregularly sampled or non-synchronous data. Such data occur in various fields including, but not limited to; financial time-series, geologic time-series, astrophysics and medical imaging [30–36]. Analysis and processing of unevenly spaced data is an ongoing field of research, as most methods for analysis assume equally spaced data.

Data used

Six proteomic MS datasets from previously published studies were used to validate the methods presented over a broad range of dataset attributes such as different numbers of peaks, different peak widths, differently spaced m/z values, different number of m/z values, different number of spectra, samples from different organisms and samples from different biological origins.

Fiedler data: Urine samples were taken from 10 healthy women and 10 healthy men and peptides were separated using magnetic beads (fractionation). The fractionated samples were then subject to MALDI TOF-MS [37]. A subset of the MALDI TOF-MS data is freely available in the R package MALDIquant [7] and is the dataset used here. The spectra are observed over the range of values 1,000-10,000 m/z.

Yildiz data: As described in [38], sera were collected from 142 lung cancer patients and 146 healthy controls to find relevant biomarkers. The serum samples were subject to MALDI TOF-MS without magnetic bead separation. The spectra are observed over the range of values 3,000-20,000 m/z.

Wu data: MALDI TOF-MS data were generated from sera, as described in [39, 40], with the aim of differentiating between 47 ovarian and 42 control subjects. The spectra are observed over the range of values 800-3,500 m/z using Reflectron mode which resolve peptide peaks into their isotopomers.

Adam data: Surface-enhanced laser desorption/ionization (SELDI) TOF-MS data from 326 serum samples from subjects classified as prostate cancer, benign hyperplasia or control [41]. While SELDI has been found to be less sensitive than MALDI, samples do not require fractionation before applying MS. The data analysed here are limited to the range 2,000-15,000 m/z as peptide signals beyond this range are sparse.

Taguchi data: The dataset available was first described in [42] but is available as a supplement for [43]. The data are 210 serum-derived MALDI TOF mass spectra from 70 subjects with non-small-cell lung cancer with the aim of predicting response to treatment. The data observed cover the 2,000-70,000 m/z range.

Mantini data: The data in this study were produced using MALDI TOF-MS from purified samples containing equine myoglobin and cytochrome C [44]. A total of 30 spectra are available in the range 5,000-22,000 m/z.

Transformation of the m/z-axis

The proposed pipeline for baseline subtraction requires a suitable transformation of the m/z-axis as the first step. In this section we investigate potential transformations, that will be assessed quantitatively for their suitability.

It has previously been suggested that peak width is roughly proportional to peak location on the TOF-axis [45, 46] and that therefore peak width is proportional to the square root of the m/z location. This was not

Stanford *et al. Proteome Science* (2016) 14:19

Page 7 of 15

in fact observed for any of the datasets analysed in the present study. Various transforms that expand the low m/z values and contract the high m/z values were then investigated. Therefore, suitable transforms are functions that are $1 - 1$, monotonically increasing and non-convex. For simplicity, only functions with analytical inverses for back-transformation were considered. Table 1 sets out the shortlist of suitable Box-Cox like transformations, t_0-t_5, that are considered appropriate for application here [47].

To illustrate the role of the transformation, Fig. 5 shows a spectrum from the Fiedler dataset on the original m/z-axis (t_0) for transformations t_1 and t_3. The effect of t_3, when compared to the original m/z-axis, is an expansion of smaller mass peak widths and the contraction of higher mass peak widths. However, visually it can be seen that higher mass peaks have larger peak widths on average even under the t_3 transformation. The t_1 transformation further shifts low m/z values across the transformed axis and contracts m/z values at the high end of m/z-axis. Potentially, the t_1 transformation creates larger peak widths for smaller m/z values than high m/z values so as to produce peak widths that decrease on average across the transformed axis. The effect of the six transformation functions, t_0-t_5, on a spectrum from each of the six datasets is available in 'Additional file 2'.

Obtaining approximate peak widths prior to baseline subtraction

Peak widths can be obtained at the peak detection step (step four of pre-processing) but such information is not generally known prior to the second pre-processing step of baseline subtraction. To determine the constant SE size to be passed over the transformed m/z-axis, peak widths need to be estimated. An algorithm to estimate peak widths from the data was created here for this purpose.

The algorithm below to estimate the peak widths within spectra takes the previously smoothed spectra on the transformed m/z-axis as the input and is performed as follows.

Table 1 The transforms, t_i, of the m/z-axis trialled to produce a roughly uniform distribution of peak widths across the t_i (m/z)-axis

Label	Transform
t_0 (x)	x
t_1 (x)	$-1000x^{-1}$
t_2 (x)	$x^{1/4}$
t_3 (x)	$\ln x$
t_4 (x)	$-1000 (\ln x)^{-1}$
t_5 (x)	$-1000x^{-1/4}$

- For each spectrum, the lower convex hull of the two-dimensional set of spectrum points is used to determine an approximate baseline for each spectrum.
- The longest segment of the lower convex hull is then halved, with the two sets of points created by this split subject to a new lower convex hull calculation.
- The newly calculated lower convex hull points for the two set of points are then added to the original set of lower convex hull points to improve the approximate baseline calculation.
- This is repeated $r - 1$ more times to produce an approximate estimated baseline.
- The approximate baseline is then removed and median intensity is then calculated for the resulting spectrum.
- Intensities above the median value are treated as points along a peak.
- The consecutive points above the median value are the estimated peak widths.

The above algorithm is crude and could not be used for reliable baseline subtraction. However, estimated peak widths are easily extracted using this method and can be used within the proposed automated baseline subtraction pipeline.

A reasonable number of lower convex hull iterations of $r = 5$ produced sensible results on the six datasets used. By specifying a value of r, this method to estimate peak widths is fully automated. It provided enough alterations to the original lower convex hull to satisfactorily remove the residual baseline on concave smoothed spectra while not applying too many alterations so as to create midpoints along the longest segments which create lower convex hulls ending a peak vertices and therefore removing them. However, a missed peak or two per spectrum is not an issue as dozens of peaks are identified per spectrum. Figure 6 depicts this process, on a single spectrum.

The algorithm presented above attempts to automatically find peak widths without user input. We outline this automated procedure in the Methods section as it is not the focus of this paper, and may be substituted with any peak region finding method that requires no user input; such is the modularity of the pipeline shown in Fig. 4. There exist other methods to estimate peak widths (regions), such as that found in [20], but they require previous knowledge of likely peak widths and are therefore not a baseline subtraction method that can be automated.

Selecting a SE size and applying the top-hat operator in the transformed space

Point three of Fig. 4 requires a choice of SE size. This can be chosen from the estimated peak widths found using
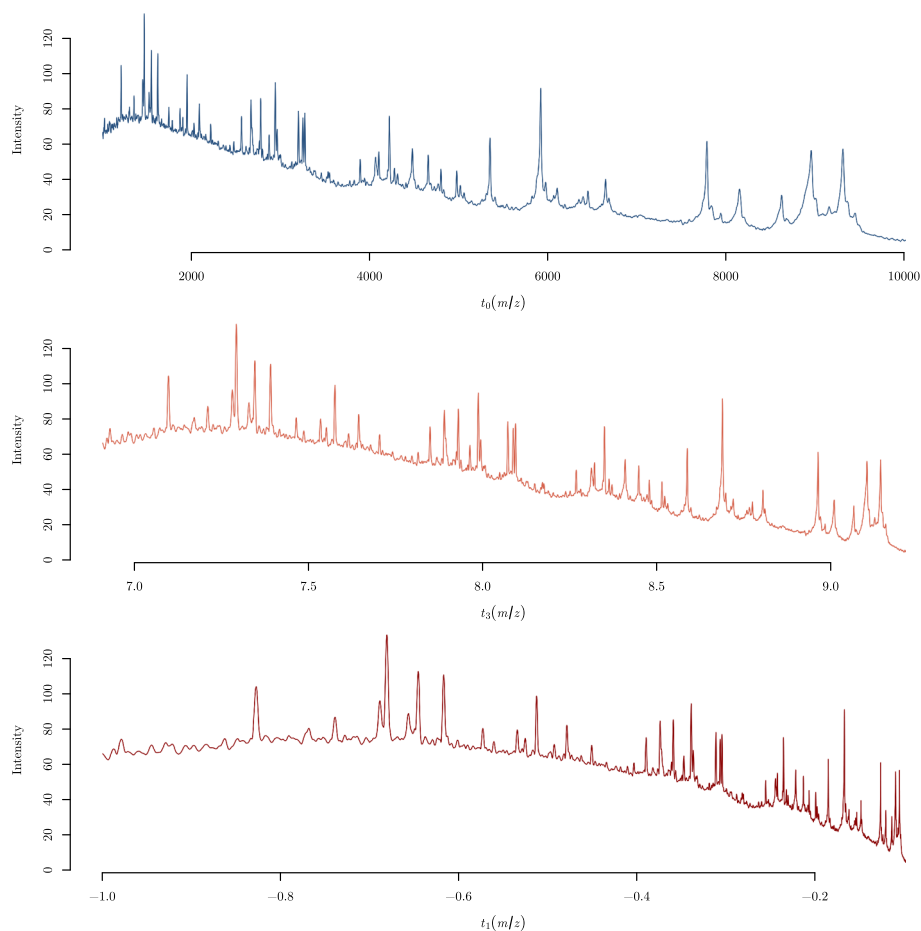
Stanford *et al. Proteome Science* (2016) 14:19

Page 8 of 15

Fig. 5 Transformations of the *m/z*-axis: Three different *m/z*-axis transformations (see Table 1) for the exemplar Fiedler spectrum given in Figs. 2 and 3

the algorithm presented in previous section. The aim is to select a SE of sufficient size to not undercut peaks; such a SE size roughly translates to the maximum of the peak widths. However, there is likely to be a SE size smaller than the maximum estimated peak width but much greater than the minimum estimated peak width that performs optimally. Given a set of estimated peak widths for all spectra in an experiment and a SE size, we define the proportion of peak widths that are estimated to be the SE size or smaller as the estimated peak coverage proportion (EPCP). Please refer to 'Additional file 3' for an illustrative plot of the estimated peak widths for the 16 spectra in the Fiedler dataset on the t_2 (*m/z*) scale.

We trial different SE sizes corresponding to different EPCP values in the hope an optimal EPCP value for each of the six datasets we utilise can be found. A SE size that fully covers 95% of detected estimated peak widths (EPCP of 0.95) for example, could yield optimised baseline subtraction.
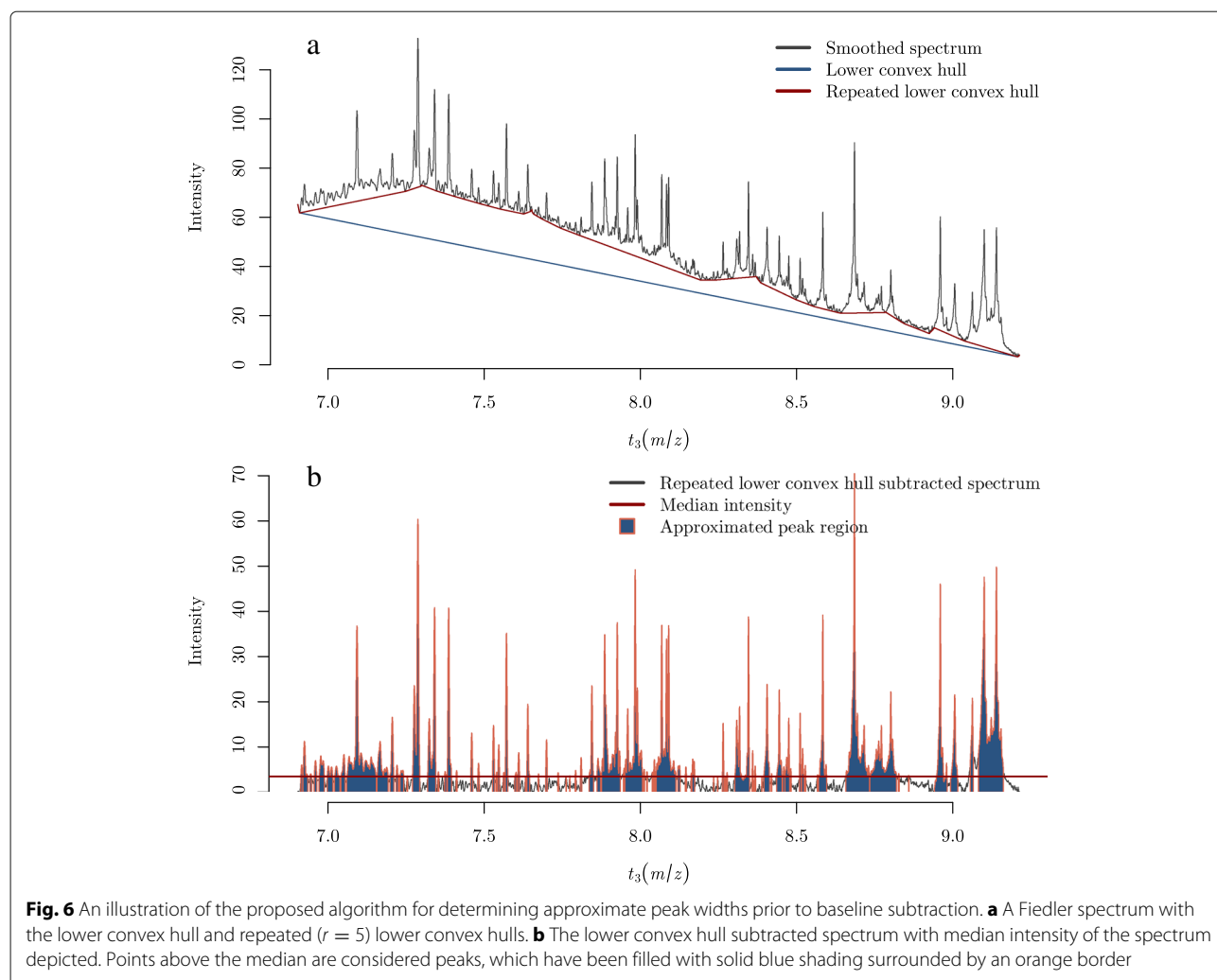
Both the EPCP and *m/z*-axis transformation are variables that can be tuned to find an empirically optimal

combination, assessed by calculating the minimum value of an error metric relative to a gold-standard baseline subtraction. The metric used to compare the automated baseline subtraction to the gold-standard is outlined in the next section and the modified algorithm to perform top-hat baseline subtraction on the unevenly spaced and transformed *m/z*-axis is provided in the section after that.

Comparison of proposed methods to the gold-standard

Piecewise, top-hat baseline subtracted spectra were used as the gold-standard baseline subtracted spectra. The SE sizes for each piecewise segment along the *m/z*-axis were selected using trial-and-error to produce the best baseline subtraction as determined visual inspection. These baseline subtracted, gold-standard spectra were produced prior to the automated baseline subtraction methods being applied.

Mean absolute scaled error (MASE [48]) was selected to be the error metric of the automatically baselined spectra

Stanford *et al. Proteome Science* (2016) 14:19

Page 9 of 15

Fig. 6 An illustration of the proposed algorithm for determining approximate peak widths prior to baseline subtraction. **a** A Fiedler spectrum with the lower convex hull and repeated ($r = 5$) lower convex hulls. **b** The lower convex hull subtracted spectrum with median intensity of the spectrum depicted. Points above the median are considered peaks, which have been filled with solid blue shading surrounded by an orange border

for a given transformation and EPCP, when compared to the gold-standard baseline subtracted spectra. Because the MALDI TOF mass spectra intensities are on arbitrary scales prior to normalisation, it is important to use a metric that is scale free, in order to be able to compare results between spectra from different experiments. MASE also avoids many degeneracy issues of other relative error metrics with zero denominators. Baseline subtracted spectra will have many zero values where no signal is present. Other metrics such as mean squared error (MSE) were considered (which did not change the selection of the optimal transform and EPCP) however the ability to compare the error with other data is not possible and some sort of normalisation or weighting of spectra is required to ensure the MSE, say, of selected spectra do not dominate the result.

Let τ_j^* denote the intensity at x_j of a gold-standard baseline subtracted spectrum $\tau_{B^*}(x_j)$ and τ_j denote an automated baseline subtracted spectrum $\tau_B(x_j)$. The MASE is calculated as

$$\text{MASE} = \text{mean}\left(\left\{\frac{\left|\tau_j^* - \tau_j\right|}{\frac{1}{n-1}\sum_{i=2}^{n}\left|\tau_i^* - \tau_{i-1}^*\right|}\right\}_{j=1,2,\ldots,n}\right).$$

For each of the six datasets, there are N spectra to be compared. Let AMASE be the average MASE value of the N baseline subtracted spectra, then

$$\text{AMASE} = \frac{1}{N}\sum_{\ell=1}^{N}\text{MASE}_\ell.$$

The 'continuous' line segment algorithm
A novel algorithm is proposed here that can be applied to the unevenly spaced values of the transformed m/z-axis using a constant SE width. This algorithm, which we name the 'continuous' line segment algorithm (CLSA), requires fewer computations per element than current rolling maximum and minimum algorithms on unevenly spaced data [49].

Consider the case where values along the x-axis are not evenly spaced, such as proteomic spectra on a

Stanford *et al. Proteome Science* (2016) 14:19

Page 10 of 15

transformed $t\,(m/z)$-axis, in contrast to a simpler case where the x-axis values are the integers $1, 2, \ldots, n$. Figure 7 outlines the CLSA as a rolling minimum algorithm that can be trivially converted to a rolling maximum algorithm by finding the rolling minimum of $-f$ and returning the negative values of the result.

In effect, the CLSA creates m blocks using the θ_i relating to the corresponding x_i:

$$
\begin{aligned}
\theta_1, \theta_2, \ldots, \theta_{b_1} &= 1 \ \ \text{where } x_1, x_2, \ldots, x_{b_1} \\
&\quad \in [x_1, x_1 + k) \\
\theta_{b_1+1}, \theta_{b_1+2}, \ldots, \theta_{b_2} &= 2 \ \ \text{where } x_{b_1+1}, x_{b_1+2}, \ldots, x_{b_2} \\
&\quad \in [x_1 + k, x_1 + 2k) \\
&\vdots \\
\theta_{b_{m-1}+1}, \theta_{b_{m-1}+2}, \ldots, \theta_{b_m} &= m \ \ \text{where } x_{b_{m-1}+1}, x_{b_{m-1}+2}, \ldots, \\
&\quad x_{b_m} \in [x_1 + (m-1)k, x_n].
\end{aligned}
$$

When the algorithm considers each point x_i for the minimum f in the window spanning $k/2$ either side, it checks whether the most extreme x-values in this window are either in the current block or one block away (these values cannot be further than one block away as block sizes are of length k) to decide on which combination of g and h is required. Note the algorithm is impervious to arbitrarily spaced x_i as long as they are in ascending order. If $\theta_i \neq j$ for any $i = 2, 3, \ldots, n-1; j = 2, 3, \ldots, m-1$ (empty blocks) or $x_{b_{j-1}+1} = x_{b_j}$ for any $j = 2, 3, \ldots, m$ (blocks with only one x_i), for example, do not affect the validity of the proposed algorithm.

This algorithm can be seen as a generalised version of the LSA [25, 26] as it works on evenly and unevenly spaced data. An R implementation of this novel CLSA can be found as an R-package using compiled C code at https://github.com/tystan/clsa.

Let $k_0 = \frac{k}{2}$ where k is the length of a centred, one-dimensional window. Consider the ordered (ascending) set of transformed m/z points, $X = \{x_1, x_2, \ldots, x_n\}$, and the corresponding expression values, f. Furthermore, define $\mathrm{span}(X) = x_n - x_1$ where $\mathrm{span}(X) > k$ and choose the smallest $m \in \mathbb{Z}^+$ so that $mk \geq \mathrm{span}(X)$.

Three vectors taking integer values are required to be created initially,

$$
\Theta = [\theta_1, \theta_2, \ldots, \theta_n], W^\nabla = \left[w_1^\nabla, w_2^\nabla, \ldots, w_n^\nabla\right], W^\triangle = \left[w_1^\triangle, w_2^\triangle, \ldots, w_n^\triangle\right].
$$

For $i = 1, 2, \ldots, n$, the integer θ_i is calculated as follows,

$$
\theta_i = \{j : \text{if } x_1 + (j-1)k \leq x_i < x_1 + jk\} \ \text{for } j \in \{1, 2, \ldots, m\};
$$

w_i^∇ is the index corresponding to x_i satisfying the inequality,

$$
x_{w_i^\nabla - 1} < x_i - k_0 \leq x_{w_i^\nabla};
$$

and $w_i^\triangle$ is the index corresponding to x_i satisfying the inequality,

$$
x_{w_i^\triangle} \leq x_i + k_0 < x_{w_i^\triangle + 1}.
$$

The rolling minimum at x_i can be calculated as,

$$
r_{\min}\left(f\left(x_i\right)\right) = \begin{cases} g\left(x_{w_i^\triangle}\right) & \text{if } \theta_{w_i^\nabla} = \theta_{w_i^\triangle + 1} \\ h\left(x_{w_i^\nabla}\right) & \text{if } \theta_{w_i^\nabla - 1} = \theta_{w_i^\triangle} \\ \min\left\{g\left(x_{w_i^\triangle}\right), h\left(x_{w_i^\nabla}\right)\right\} & \text{otherwise,} \end{cases}
$$

where

$$
g\left(x_i\right) = \begin{cases} f\left(x_i\right) & \text{if } \theta_{i-1} < \theta_i \ (\text{define } \theta_0 = 0) \\ \min\left\{g\left(x_{i-1}\right), f\left(x_i\right)\right\} & \text{otherwise; and} \end{cases}
$$

$$
h\left(x_i\right) = \begin{cases} f\left(x_i\right) & \text{if } \theta_i < \theta_{i+1} \ (\text{define } \theta_{n+1} = m + 1) \\ \min\left\{f\left(x_i\right), h\left(x_{i+1}\right)\right\} & \text{otherwise.} \end{cases}
$$

Fig. 7 The continuous line segment algorithm (CLSA)

Stanford *et al. Proteome Science* (2016) 14:19

Page 11 of 15

A demonstration of why the creation of blocks the size of the SE and accessing cumulative values half an SE length away allows the calculation of rolling minimums is shown in [25]. Examples to demonstrate the mechanics of the CLSA algorithm are presented in 'Additional file 4'.

Results and discussion

Presented in this paper is a pipeline to automate the baseline subtraction step in proteomic TOF-MS pre-processing. The pipeline consists of transforming the m/z-axis, then finding an appropriate SE size via an automated peak width estimation algorithm on the transformed scale, applying a novel algorithm to perform the top-hat baseline subtraction, then finally, baseline subtracted spectra are returned by back-transforming the data to the m/z scale.

There remain two elements of the pipeline to be assessed. Firstly, for the pipeline to be fully automated, an optimal combination of EPCP value and transformation need to be found. In the next section we perform a grid search over EPCP values of $0.8, 0.85, 0.9, 0.95, 0.98, 0.99, 1$ and transformations $t_0, t_1, t_2, t_3, t_4, t_5$ to find which combination provides the closest baseline subtracted signal to the gold-standard. Given sufficient similarity to the gold-standard is achieved, it is hoped that a consensus over all datasets, in their varying attributes, of the optimal combination of EPCP value and transformation can be found. If a consensus is indeed found, the pipeline is likely to be applicable to other proteomic TOF-MS datasets.

A theoretical and empirical assessment of the efficiency of the CLSA in comparison to naive rolling window algorithms then follows. The theoretical efficiency is discussed with respect to the number of operations required over all the elements input into the CLSA. By performing the top-hat operation on the six proteomic TOF-MS datasets and simulated datasets of varying sizes, the computational time required for the CLSA versus the naive algorithm provides an empirical assessment of their relative efficiencies.

Comparison of piecewise and transformed axis baseline subtraction

No single transformation or EPCP was optimal, however, EPCP between 0.95 and 0.99 provided the optimal AMASE value for all datasets suggesting the peak width estimation process is relatively stable. On the Fiedler, Yildiz, Taguchi and Mantini datasets, the null transformation which implicitly implies a constant peak width across the m/z-axis was not valid as AMASE values were notably higher than for the remaining transformations. The transformations t_2, t_3, t_4 and t_5 produced the best results. It should be noted that the transformations t_3, t_4 and t_5 produced very similar AMASE values. With the exception of the Yildiz dataset, using these transformations with an EPCP of 0.95 produced sensible results. Please see 'Additional file 3' for more details relating to the transformation and EPCP optimisation.

Figure 8 demonstrates the baseline estimates using the gold-standard piecewise top-hat operator, the AMASE optimal transformation and EPCP (t_3, 0.98) and a non-optimal combination of transformation and EPCP (t_4, 0.95) that was suitable on all but the Yildiz data. The
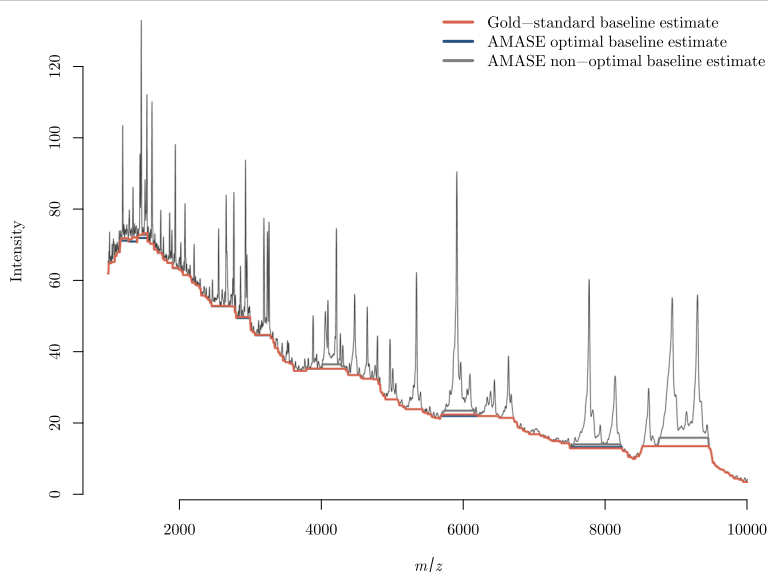

Fig. 8 Optimised automated baseline estimate (*blue*) compared to the gold-standard (*orange*) piecewise baseline estimate for the Fiedler spectrum. The optimal transformation and EPCP were t_3 and 0.98, respectively. A non-optimal transformation (*grey*) is provided for reference, for which the corresponding transformation and EPCP are t_4 and 0.95, respectively

Stanford *et al. Proteome Science*   (2016) 14:19

Page 12 of 15

optimal AMASE transformation and EPCP combination (t_3, 0.98) shows very little difference from the gold-standard baseline estimate.

Because the gold-standard baseline estimate is subject to expert input and opinion, the differences seen in the gold-standard and the optimal AMASE baseline estimate are not of concern as both look sensible. The non-optimal baseline estimate produces a reasonable automated baseline subtraction, however, it can be seen that this estimate does undercut the peaks especially at high m/z-values.

Efficiency of the CLSA compared to the naive rolling window

The naive rolling minimum algorithm consists of the linear-time process of finding the indexes of points at the upper and lower edges of the sliding window for each element, by incrementing the edge indexes from the previous element when required. Using a_k as the average number of data points in the sliding window of size k, the computational cost of finding the minimum value in the window requires approximately $a_k - 1$ comparisons per element. This is because each element requires, on average, a minimum or maximum comparison of all the data points in the window except one: the first data point does not require a comparison. The resulting computational complexity is $\mathcal{O}(a_k n)$ for the naive algorithm, which is dependent on the size of the sliding window and the number of elements in X.

Like the LSA, the CLSA is a linear-time algorithm irrespective of the window size, k. For the CLSA, a linear-time progression through the n elements is required to assign integers of the Θ-vector, as each element is an integer equal to or greater than that which precedes it. The linear-time process of finding the $W^\triangledown$ and $W^\triangle$ indexes at the lower and upper edges of the sliding window, respectively, for each element is similar to that required in the naive algorithm. One linear-time sweep forward and one linear-time sweep back on the data is required to create g and h. A final sweep of the created vectors $W^\triangledown$, $W^\triangle$, Θ, g and h is required to compute the $r_{\min}$ values. Each $r_{\min}\left(f\left(x_i\right)\right)$ calculation requires the tests $\theta_{w_i^\triangledown} = \theta_{w_i^\triangledown + 1}$, $\theta_{w_i^\triangledown - 1} = \theta_{w_i^\triangle}$ or $\min\left\{g\left(x_i\right), h\left(x_i\right)\right\}$. It can therefore be deduced the CLSA is $\mathcal{O}(n)$ complexity, requiring a series of linear-time operations, importantly independent of the length of the sliding window, k.

Given the MS application, $a_k - 1$ operations per element in the naive algorithm would be much larger than the constant number of operations required per element for the CLSA and efficiency strongly favours the CLSA. It should be pointed out that the CLSA requires extra memory availability beyond the iterative algorithm for the creation of the vectors $W^\triangledown$, $W^\triangle$, Θ, g and h. Another computational advantage of the CLSA is that by using the minimum of the two temporary vectors g and h as

opposed to the minimum of a non-constant number of data points for each $x_i \in X$, vectorised programming can be utilised instead of loops. This is of significant advantage in programming languages that are interpreted such as R.

Using the clsa package, the CLSA and naive sliding window algorithms were compared for computational time to calculate the top-hat on real and simulated data. The computations were performed on a 21.5" iMac (late 2013 model, 2.7GHz Intel Core i5, 8GB 1600MHz DDR3 memory, OS X 10.10.2). To optimise speed, the calculations requiring iterative looping were performed using compiled C code for both the CLSA and naive algorithms. The code to run the test of computational running time on the simulated data is provided in 'Additional file 5'.

The CLSA and naive sliding window algorithms were applied to perform top-hat baseline estimation to the six datasets used in this paper and the results are shown in Table 2. The CLSA resulted in a reduction of the required computational time by a factor of at least 4. The advantage in speed of the CLSA had greater improvement for the datasets with a greater number of m/z values. The biggest relative improvement was by a factor exceeding 50 for the largest dataset in terms of m/z values per spectra on the Yildiz spectra.

Table 3 displays the computational times of top-hat baseline estimation using the CLSA and naive algorithms for varying datasets and SE sizes. The simulated data consisted of 20 randomly generated spectra with x_i and f_i for $i = 1, 2, \ldots, n$. These values were independently and randomly generated, where the signal locations $x_i \sim$ Beta $(1, 3)$ mimic a higher density of points at the low end of the spectra and $f_i \sim \chi_{10}^2$ mimic the positive signals in spectra. MALDI TOF-MS data can have in excess of tens of thousands of m/z values, hence, values of $n = 10^4, 2 \times 10^4, \ldots, 10^5$ were used. Varying window sizes were tested, ranging in width from 0.5 to 20% of the x-axis domain. i.e., 0.5% corresponds to a window size of 0.005 passed over the domain $[0, 1]$.

Table 2 Computational time to perform top-hat baseline subtraction in the transformed space using the naive and CLSA algorithms on the six datasets under study

Data	Number of specta	Number of m/z values	Computational time (sec)	
			Naive algorithm	CLSA
Fiedler	16	42388	7.7	0.2
Yildiz	264	75958	312.6	5.5
Wu	89	91378	34.1	1.7
Adam	326	8461	3.0	0.7
Taguchi	210	19234	18.0	0.9
Mantini	30	32967	6.7	0.2

Stanford *et al. Proteome Science* (2016) 14:19

Page 13 of 15

Table 3 Computational time in seconds to perform top-hat baseline subtraction in the transformed space using the naive and CLSA algorithms on synthetic data for varying data assumptions and SE sizes

Number of points n ($\times 10^4$)	Naive Window size (% of x-axis)						CLSA Window size (% of x-axis)					
	0.5	1	2	5	10	20	0.5	1	2	5	10	20
1	0.1	0.1	0.3	0.6	1.2	2.3	0.0	0.0	0.2	0.0	0.0	0.2
2	0.3	0.5	1.0	2.5	4.9	9.2	0.1	0.1	0.1	0.1	0.1	0.1
3	0.6	1.2	2.3	5.7	11.0	20.6	0.3	0.1	0.1	0.2	0.1	0.1
4	1.1	2.1	4.2	10.2	19.6	36.5	0.1	0.2	0.1	0.1	0.3	0.1
5	1.7	3.3	6.5	15.8	30.5	56.8	0.2	0.3	0.2	0.2	0.3	0.2
6	2.4	4.7	9.3	22.7	44.0	82.0	0.2	0.2	0.3	0.2	0.2	0.3
7	3.2	6.4	12.6	30.9	59.7	111.7	0.2	0.2	0.4	0.2	0.2	0.4
8	4.2	8.4	16.6	40.3	78.3	146.4	0.4	0.3	0.3	0.4	0.3	0.3
9	5.4	10.6	21.0	51.1	98.8	185.3	0.4	0.3	0.3	0.4	0.3	0.3
10	6.6	13.0	25.9	63.1	121.8	228.7	0.3	0.4	0.3	0.3	0.5	0.3

The CLSA was faster than the naive algorithm in every scenario as shown in Table 3. As expected, the computational time was constant for the CLSA irrespective of the window size for a fixed number of points (number of transformed m/z values). The difference in computational time between the two algorithms was reasonably small for small datasets and small SE sizes. However, for a typical number of m/z points seen in practice, say 50,000, and a moderate window size that on average encapsulates 5,000 points (1% of x-axis), the CLSA provides an order of magnitude increase in speed.

Conclusion

The current gold-standard in baseline subtraction is a piecewise approach that is performed manually, that is, by inspection. Piecewise baseline subtraction is typically performed because, as we have consistently observed with the datasets analysed in this paper, the properties of the spectra do not remain constant over their domain. In particular, a spectrum's peak width increases with increasing m/z-values. We have proposed a new baseline subtraction pipeline be adopted for the correction of mass proteomic spectra data which avoids both the manual user input and the piecewise-subtraction aspect of existing methods. Our new pipeline is based on the premise that a suitable transformation of the m/z-axis can be found which removes the relationship between peak width and peak location.

As part of the new pipeline, we propose a method to create data-based, and therefore data specific, peak-width estimates from smoothed spectra. Even if this step is not used to automate baseline subtraction, it provides an initial sensible SE size that adapts to each individual dataset. Our generalised version of the LSA is also presented in the paper, which we call CLSA. CLSA can be applied to unevenly or evenly spaced data and is not limited in its application to proteomic MS data. Should a transformation be known to create peak widths independent of m/z-location in proteomic MS data, an efficient and effective baseline subtraction can be performed using the top-hat operator with a CLSA implementation. A major contribution to note is that we have demonstrated CLSA far outperforms the naive rolling minimum algorithm in required computational time by an order of magnitude or more on numerous datasets of real-world complexity.

The transformed and constant-sized window approach may suffer from a slight but largely unnoticeable reduction in sensitivity in comparison. The trade-off between exactness of the piecewise approach and the speed of the automated transformation and continuous approach may be a consideration, especially if a known relationship exists between the peak width and peak location.

Availability of supporting data

Fiedler data: A subset of the MALDI TOF-MS data generated by the study [37] is available in the publicly available R package: MALDIquant [7].

Yildiz data: Available at http://www.vicc.org/biostatistics/serum/JTO2007.htm.

Wu data: Previously available at http://bioinformatics.med.yale.edu/MSDATA.

Adam data: Data was obtained on request from the authors of [41]. However some Eastern Virginia Medical School data is available at http://edrn.nci.nih.gov/science-data.

Taguchi data: Available at http://www.vicc.org/biostatistics/download/WSData.zip.

Mantini data: Available at http://www.biomedcentral.com/content/supplementary/1471-2105-8-101-S2.zip.

Stanford *et al. Proteome Science*   (2016) 14:19

Page 14 of 15

Additional files

Additional file 1: Morphological image analysis. An overview of the mathematical machinery used to define the top-hat operator. A small example is provided to demonstrate how the top-hat operator is applied to data. (PDF 240 kb)

Additional file 2: Six transformation functions on each of the six datasets. The effect of the six transformation functions, t_0-t_5, on a spectrum from each of the six datasets. (PDF 4669 kb)

Additional file 3: EPCP and m/z-axis transformation optimisation. The AMASE results on each of the six datasets for varying EPCP and m/z-axis transformations as well as an illustrative plot of estimated peak widths on the t_2(m/z) scale. (PDF 633 kb)

Additional file 4: CLSA examples. Examples with calculations to demonstrate the continuous line segment algorithm. (PDF 308 kb)

Additional file 5: CLSA and naive algorithm computational times for simulated data. R code to produce the top-hat baseline subtraction computational time results shown in Table 3. (PDF 30.9 kb)

Abbreviations
AMASE: Average mean absolute scaled error; CLSA: Continuous line segment algorithm; Da: Dalton; EPCP: Estimated peak coverage proportion; LOESS: Locally weighted scatterplot smoothing; LSA: Line segment algorithm; MALDI: Matrix-assisted laser desorption/ionization; MASE: Mean absolute scaled error; MS: Mass spectrometry; MSE: Mean squared error; SNIP: Sensitive nonlinear iterative peak; TOF: Time of flight

Authors' contributions
TS and PS developed the statistical and analytical methods. CB and PS provided guidance on the analysis of proteomic data. TS developed the code and implementation. All authors contributed to the writing of the manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
The research was exempt from formal University of Adelaide Human Research Ethics Committee approval according to the Australian National Statement on Ethical Conduct in Human Research, 2007.

References
1. Albrethsen J. Reproducibility in protein profiling by MALDI-TOF mass spectrometry. Clin Chem. 2007;53(5):852–8.
2. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nat Clin Pract Oncol. 2008;5(10):588–99.
3. Hortin GL. The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome. Clin Chem. 2006;52(7):1223–37.
4. Croxatto A, Prod'hom G, Greub G. Applications of maldi-tof mass spectrometry in clinical diagnostic microbiology. FEMS Microbiol Rev. 2012;36(2):380–407.
5. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2014. http://www.R-project.org/.
6. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 2004;5:80.
7. Gibb S, Strimmer K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. Bioinformatics. 2012;28(17):2270–1.
8. Li X. PROcess: Ciphergen SELDI-TOF Processing. 2005. R package version 1.42.0. http://bioconductor.org/packages/release/bioc/html/PROcess.html. Accessed July 2015.
9. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. Anal Chem. 2006;78:779–87.
10. Stanford TE. Statistical analysis of proteomic mass spectrometry data for the identification of biomarkers and disease diagnosis. PhD thesis, School of Mathematical Sciences, The University of Adelaide;. 2015.
11. Glish GL, Vachet RW. The basics of mass spectrometry in the twentyfirst century. Nat Rev Drug Discov. 2003;2(2):140–50.
12. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem. 1964;36(8):1627–39.
13. Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. J Comput Graph Stat. 2002;11: 108–36.
14. Mayer CD, Glasbey CA. Statistical methods in microarray gene expression data analysis. In: Husmeier D, Dybowski R, Roberts S, editors. Probabilistic Modeling in Bioinformatics and Medical Informatics. Advanced Information and Knowledge Processing. London: Springer; 2005. p. 211–38.
15. Sauve AC, Speed TP. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. In: Proceedings of the Genomic Signal Processing and Statistics workshop. John Hopkins University, Baltimore, MD, May 26–27; 2004.
16. Kohlbacher O, Reinert K, Gröpl C, Lange E, Pfeifer N, Schulz-Trieglaff O, Sturm M. TOPP-the OpenMS proteomics pipeline. Bioinformatics. 2007;23(2):191–7.
17. Lange E, Gröpl C, Schulz-Trieglaff O, Leinenbach A, Huber C, Reinert K. A geometric approach for the alignment of liquid chromatography-mass spectrometry data. Bioinformatics. 2007;23(13):273–81.
18. Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, Kohlbacher O. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics. 2008;9(1):163.
19. Bauer C, Kleinjung F, Smith C, Towers M, Tiss A, Chadt A, Dreja T, Beule D, Al-Hasani H, Reinert K, Schuchhardt J, Cramer R. Biomarker discovery and redundancy reduction towards classification using a multi-factorial maldi-tof ms t2dm mouse model dataset. BMC Bioinformatics. 2011;12(1):140.
20. Morháč M. An algorithm for determination of peak regions and baseline elimination in spectroscopic data. Nuclear Instruments Methods Phys Res Sect A Accelerators Spectrometers Detectors Assoc Equip. 2009;600(2): 478–87.
21. Ryan CG, Clayton E, Griffin WL, Sie SH, Cousens DR. SNIP, a statistics-sensitive background treatment for the quantitative analysis of {PIXE} spectra in geoscience applications. Nuclear Instruments Methods Phys Res Sect B Beam Interact Mater Atoms. 1988;34(3):396–402.
22. Yang C, He Z, Yu W. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. BMC Bioinformatics. 2009;10(1):4. doi:10.1186/1471-2105-10-4.
23. Dougherty E. Mathematical Morphology in Image Processing. New York: Marcel-Dekker; 1992.
24. Soille P. Morphological Image Analysis: Principles and Applications. Secaucus: Springer; 1999.
25. van Herk M. A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. Pattern Recogn Lett. 1992;13(7): 517–21.
26. Gil J, Werman M. Computing 2-D min, median, and max filters. IEEE Trans Pattern Anal Mach Intell. 1993;15:504–7.

Stanford *et al. Proteome Science* (2016) 14:19

Page 15 of 15

27. van Herk M, de Munck JC, Lebesque JV, Muller S, Rasch C, Touw A. Automatic registration of pelvic computed tomography data and magnetic resonance scans including a full circle method for quantitative accuracy evaluation. Med Phys. 1998;25:2054.

28. Heneghan C, Flynn J, O'Keefe M, Cahill M. Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. Med Image Anal. 2002;6(4):407–29.

29. Zhang G, Ueberheide BM, Waldemarson S, Myung S, Molloy K, Eriksson J, Chait BT, Neubert TA, Fenyö D. Protein quantitation using mass spectrometry. In: Fenyö D, editor. Computational Biology. Methods in Molecular Biology. Vol. 673. New York: Humana Press; 2010. p. 211–22.

30. Greengard L, Lee JY. Accelerating the nonuniform fast fourier transform. SIAM Rev. 2004;46(3):443–54.

31. Lo AW, MacKinlay AC. An econometric analysis of nonsynchronous trading. J Econ. 1990;45(1-2):181–11.

32. Aris A, Shneiderman B, Plaisant C, Shmueli G, Jank W. Representing unevenly-spaced time series data for visualization and interactive exploration. In: Human-Computer Interaction-INTERACT 2005. Heidelberg: Springer; 2005. p. 835–46.

33. Schulz M, Mudelsee M. REDFIT: estimating red-noise spectra directly from unevenly spaced paleoclimatic time series. Comput Geosci. 2002;28(3): 421–6.

34. Deeming TJ. Fourier analysis with unequally-spaced data. Astrophys Space Sci. 1975;36(1):137–58.

35. Scargle JD. Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. Astrophys J. 1982;263:835–53.

36. Bourgeois M, Wajer F, van Ormondt D, Graveron-Demilly D. Modern Sampling Theory. Applied and Numerical Harmonic Analysis. In: Benedetto JJ, Ferreira PJSG, editors. Boston: Birkhäuser; 2001. p. 343–63.

37. Fiedler GM, Baumann S, Leichtle A, Oltmann A, Kase J, Thiery J, Ceglarek U. Standardized peptidome profiling of human urine by magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clin Chem. 2007;53(3):421–8.

38. Yildiz PB, Shyr Y, Rahman JSM, Wardwell NR, Zimmerman LJ, Shakhtour B, Gray WH, Chen S, Li M, Roder H, Liebler DC, Bigbee WL, Siegfried JM, Weissfeld JL, Gonzalez AL, Ninan M, Johnson DH, Carbone DP, Caprioli RM, Massion PP. Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. J Thoracic Oncol Off Publ Intl Assoc Study Lung Cancer. 2007;2(10):893.

39. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics. 2003;19(13): 1636–43.

40. Yu W, Li X, Liu J, Wu B, Williams KR, Zhao H. Multiple peak alignment in sequential data analysis: a scale-space-based approach. IEEE/ACM Trans Comput Biol Bioinform (TCBB). 2006;3(3):208–19.

41. Adam BL, Qu Y, Davis JW, Ward MD, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Jr GLW, Clements MA, Cazares LH. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res. 2002;62(13):3609–14.

42. Taguchi F, Solomon B, Gregorc V, Roder H, Gray R, Kasahara K, Nishio M, Brahmer J, Spreafico A, Ludovini V, Massion PP, Dziadziuszko R, Schiller J, Grigorieva J, Tsypin M, Hunsucker SW, Caprioli R, Duncan MW, Hirsch FR, Bunn PA, Carbone DP. Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. J Natl Cancer Inst. 2007;99(11):838–46.

43. Li M, Chen S, Zhang J, Chen H, Shyr Y. Wave-spec: a preprocessing package for mass spectrometry data. Bioinformatics. 2011;27(5):739–40.

44. Mantini D, Petrucci F, Pieragostino D, Del Boccio P, Di Nicola M, Di Ilio C, Federici G, Sacchetta P, Comani S, Urbani A. LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. BMC Bioinformatics. 2007;8(1):101.

45. Siuzdak G. The expanding role of mass spectrometry in biotechnology. San Diego: MCC Press; 2006.

46. House LL, Clyde MA, Wolpert RL. Bayesian nonparametric models for peak identification in MALDI-TOF mass spectroscopy. Ann Appl Stat. 2011;5(2B):1488–511.

47. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Series B (Methodological). 1964;26(2):211–52.

48. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Intl J Forecasting. 2006;22(4):679–88.

49. Eckner A. Algorithms for unevenly-spaced time series: Moving averages and other rolling operators. Technical report, Working Paper. 2013. http://www.eckner.com/papers/ts_alg.pdf. Accessed June 2015.