BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 19 2013, pages 2445–2451
doi:10.1093/bioinformatics/btt376

Data and text mining

Advance Access publication July 3, 2013

Statistical agglomeration: peak summarization for direct infusion lipidomics

Rob Smith[1,*], Tamil S. Anthonymuthu[2], Dan Ventura[1] and John T. Prince[2,*]

[1]Department of Computer Science and [2]Department of Chemistry, Brigham Young University, Provo, UT 84602, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Quantification of lipids is a primary goal in lipidomics. In direct infusion/injection (or shotgun) lipidomics, accurate downstream identification and quantitation requires accurate summarization of repetitive peak measurements. Imprecise peak summarization multiplies downstream error by propagating into species identification and intensity estimation. To our knowledge, this is the first analysis of direct infusion peak summarization in the literature.

Results: We present two novel peak summarization algorithms for direct infusion samples and compare them with an off-machine *ad hoc* summarization algorithm as well as with the propriety Xcalibur algorithm. Our statistical agglomeration algorithm reduces peakwise error by 38% mass/charge (m/z) and 44% (intensity) compared with the *ad hoc* method over three datasets. Pointwise error is reduced by 23% (m/z). Compared with Xcalibur, our statistical agglomeration algorithm produces 68% less m/z error and 51% less intensity error on average on two comparable datasets.

Availability: The source code for Statistical Agglomeration and the datasets used are freely available for non-commercial purposes at https://github.com/optimusmoose/statistical_agglomeration. Modified Bin Aggolmeration is freely available in MSpire, an open source mass spectrometry package at https://github.com/princelab/mspire/.

Contact: 2robsmith@gmail.com or jtprince@chem.byu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 17, 2012; revised on May 17, 2013; accepted on June 26, 2013

1 INTRODUCTION

Direct infusion (injection) lipidomics, sometimes called 'shotgun' lipidomics for its similarity to shotgun genomics, is an emerging but a well-studied field (Ejsing *et al.*, 2006; Ekroos *et al.*, 2002; Watson, 2006). Here, a liquid sample is injected into a mass spectrometer, yielding a set of [mass/charge (m/z), intensity, retention time (RT)] 3-tuples (Han and Gross, 2005). For our purposes, we define a data point as a single m/z and intensity observation of a given isotope at a particular RT and a peak as the data points that comprise the observation of a distinct isotope. (Hereafter, we will more accurately use the term *ridge* instead of *peak* because of the fact that direct injection lipid intensity does not vary as a function of time.) Because there is no chromatographic separation in direct infusion lipidomics, each RT scan represents an independent measurement of the sample. Ideally, the species in the sample would be uniformly distributed across RT and measured in near-identical intensities across RT, making reduction to a single 2D vector of unique ridges trivial. Unfortunately, there are several noise factors that appear in real-world direct infusion samples. Sample distribution heterogeneity results in inter-scan variance in both m/z and intensity. What is more, technical and mechanical limitations in the mass spectrometer inculcate even more error into the output. Accurately estimating the true ridge values from the resulting output file is a non-trivial challenge (Fig. 1).

To identify and quantify each lipid, its component ridges must somehow be isolated one from another, and the additive noise ridges removed. We will call this process *ridge summarization*. Only after ridge summarization can the isotopic envelopes be compared with theoretical databases to identify and quantify the individual lipids in the sample.

The necessity of a solution for the ridge summarization problem in every direct infusion lipidomics application and the presumed effect of the results of such a solution on downstream quantitation would suggest that a description of ridge summarization be found in every shotgun lipidomics study (Samuelsson *et al.*, 2004). However, it is frequently left unmentioned (e.g. Ejsing *et al.*, 2006; Orešič, 2009; Song *et al.*, 2007). Although direct infusion methods have been around since the mid-1990s, we are only aware of two published solutions to this segment of the quantitation pipeline. The first is that of treating a survey scan as a true ridge measurement (Schwudke *et al.*, 2006). From a glance at a typical shotgun lipidomics plot, it should be clear that treating any single RT scan of data as a representative set of true ridges would be less than ideal, as the scan would include many ridges with incorrect m/z and intensity and exclude many other true ridges (Fig. 1). The second, a more robust approach, applies to shotgun lipidomics a technique that has been used in several proteomics studies (Frank *et al.*, 2008; Liu *et al.*, 2007). This approach, which we label the fixed-width algorithm, averages scans across the RT dimension to yield an estimation of the true contents of the sample (Herzog *et al.*, 2011). Though this approach is simple to code and runs in linear time, it is non-statistical and does not take into account the data densities along the m/z axis.

Here, we present two statistical approaches to solving the ridge summarization problem and evaluate them against both synthetic and real-world ridge summarization problems. We also provide the first comparative performance analysis of Xcalibur and the fixed-width algorithm on the ridge summarization problem.

*To whom correspondence should be addressed.

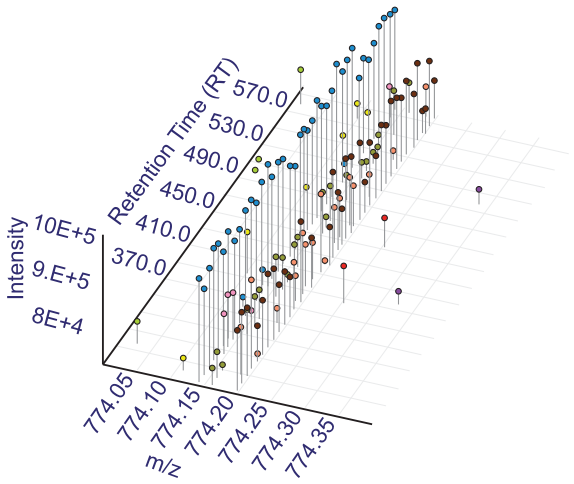

Fig. 1. A typical direct infusion lipidomics sample. The lack of consistent repetition in data points in the RT dimension and the abundance of noise in each of the three dimensions make accurate ridge summarization difficult. The colors delineate observed ridges. See online version for color

2 SYSTEM AND METHODS

We use a representative sample of three labeled datasets to test the capabilities of the methods we present, as well as the baseline results from the widely used Xcalibur software shipped with Thermo mass spectrometers.

2.1 Data

The methods presented in this article were evaluated on one synthetic dataset and two real-world hand-labeled datasets.

The Noyce dataset is a synthetic dataset constructed as described in a study by Noyce *et al.* (2013) with sampling rate 1, noise density factor 500 and 1 dimension mode.

Sample_3_750-800 and Sample_3_1000-1050 are two m/z intervals of a rat soleus lipid extract (Supplementary Material for experimental protocol). Each ridge in these datasets was isolated and labeled by hand using TOPPView (Sturm and Kohlbacher, 2009) and an exhaustive list of all (m/z, intensity, RT) triplets in the file.

Each of the datasets used in lipidomics can be represented as a list of points where each point is a (m/z, intensity, RT) triplet. For the purposes of the algorithms detailed here, points are reduced to m/z and intensity values (Fig. 2).

2.2 Metrics

Each of the following metrics measures a different quality of ridge assignment. Because each algorithm has different strengths, these metrics allow a ranking of algorithms based on what is important for the practitioner. Because we cast the ridge selection problem as a clustering problem, all of the following metrics are established clustering metrics, with the exception of normalized true ridge distance, which is a metric devised specifically for measuring the quality of summarized ridges.

In what follows, we define $\mathbf{R}$ as the set of observed ridges, $\hat{\mathbf{R}}$ as the set of predicted ridges and $\mathbf{D}$ as the set of data points.

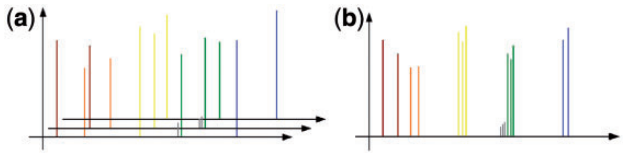

Fig. 2. Scan combination. Here (**a**) multiple scans are combined into (**b**) one list of (m/z, intensity) pairs by removing the retention time (RT) dimension. Each data point (an m/z and intensity observation of a distinct isotope at a given RT) is depicted here with a pinhead, and the collection of pinheads of one color denotes a ridge. See online version for color

We define the intensity, $\hat{r}_{int}$, of a predicted ridge $\hat{r}$ as the sum of the intensities of the ridge's assigned points:

$$\hat{r}_{int} = \sum_{d \in \hat{r}} d_{int} \tag{1}$$

and the m/z value, $\hat{r}_{m/z}$, of $\hat{r}$ as the intensity-weighted mean of the m/z value of the ridge's assigned points:

$$\hat{r}_{m/z} = \sum_{d \in \hat{r}} d_{m/z} \frac{d_{int}}{\hat{r}_{int}} \tag{2}$$

Normalized true peak distance. Normalized true peak distance (NTPD) is a metric we developed for this task, which indicates the normalized m/z or intensity difference between the predicted ridges and the nearest observed ridges. The nearest observed ridge, $\tilde{r}$, to a predicted ridge $\hat{r}$ is always calculated using m/z/ value as:

$$\tilde{r} = \underset{r \in \mathbf{R}}{\operatorname{argmin}}(|\hat{r}_{m/z} - r_{m/z}|) \tag{3}$$

Using this closest observed ridge, the m/z NTPD is calculated as:

$$\mathrm{NTPD}(\hat{\mathbf{R}}, \mathbf{R}) = \frac{1}{\min(|\hat{\mathbf{R}}|, |\mathbf{R}|)} \sum_{\hat{r} \in \hat{\mathbf{R}}} (|\hat{r}_{m/z} - \tilde{r}_{m/z}|) \tag{4}$$

and the intensity NTPD is calculated using the same equation [Equation (4)] with $\hat{r}_{m/z}$ and $r_{m/z}$ replaced with $\hat{r}_{int}$ and r_{int}.

The normalizing term controls score inflation whether the error is in predicting too many or too few ridges. The significance of this metric is reflected in its analytical relevancy. This per-ridge metric basically measures how easy it would be to correctly assign the true species label using a standard lipid species library. Such is not the case for a per-point error measure such as sum-squared error (SSE) or an intrinsic cluster metric like normalized mutual information (NMI) or purity.

Δ Number of ridges. In downstream algorithms, each estimated ridge will be treated as an actual isotope. It is clear that any identification or quantitation algorithms will be highly sensitive to the number of predicted ridges versus the number of actual ridges.

Purity. Purity measures the averaged homogeneity of each estimated ridge over all data points. It is defined as:

$$\mathrm{purity}(\hat{\mathbf{R}}, \mathbf{R}) = \frac{1}{|\mathbf{D}|} \sum_{\hat{r} \in \hat{\mathbf{R}}} \max_{r \in \mathbf{R}} |\hat{r} \cap r| \tag{5}$$

A purity of 1 is perfect, and 0 is the lowest possible score. One way to achieve high purity is to reduce the size of the predicted ridges. A naïve algorithm that simply assigns each data point into its own ridge will achieve a perfect score for purity.

Normalized mutual information. NMI allows the quantitation of the trade-off between number of predicted ridges and the quality of predicted ridges.

$$\mathrm{NMI}(\hat{\mathbf{R}}, \mathbf{R}) = \frac{\mathrm{I}(\hat{\mathbf{R}}, \mathbf{R})}{[\mathrm{H}(\hat{\mathbf{R}}) + \mathrm{H}(\mathbf{R})]/2} \tag{6}$$

where I is mutual information, given by

$$\mathrm{I}(\hat{\mathbf{R}}, \mathbf{R}) = \sum_{\hat{r}\in\hat{\mathbf{R}}} \sum_{r\in\mathbf{R}} \frac{|\hat{r} \cap r|}{|\mathbf{D}|} \log \frac{|\mathbf{D}||\hat{r} \cap r|}{|\hat{r}||r|} \tag{7}$$

and H is entropy, given by

$$\mathrm{H}(\hat{\mathbf{R}}) = -\sum_{\hat{r}\in\hat{\mathbf{R}}} \frac{|\hat{r}|}{|\mathbf{D}|} \log \frac{|\hat{r}|}{|\mathbf{D}|} \tag{8}$$

NMI indicates the dependence of sets $\hat{\mathbf{R}}$ and $\mathbf{R}$. If they are completely independent, the ridge predictions provide no information about the observed ridge assignments (indicated by an NMI of 0). A perfect score of 1 indicates that the observed ridge assignments provide no additional information beyond that provided by the predicted ridge assignments.

Sum-squared error. SSE is a common measurement of error. It is computed by summing the squared error of each assignment.

For the SSE of the m/z dimension, we use:

$$\mathrm{SSE}(\hat{\mathbf{R}}, \mathbf{R}) = \sum_{d\in\mathbf{D}} (\hat{r}^d_{m/z} - r^d_{m/z})^2 \tag{9}$$

where $\hat{r}^d_{m/z}$ indicates the m/z of the predicted ridge containing point d and $r^d_{m/z}$ indicates the m/z of the observed ridge containing point d.

Intensity SSE is calculated in the same fashion, with intensity replacing m/z in Equation (9).

3 ALGORITHMS

While both methods proposed as well as the fixed-width method follow the ridge summarization paradigm by combining multiple scans (Fig. 2), each of the three methods diverges in the way the ridges are segmented once combined into one spectrum.

3.1 Fixed ridge width method

Many practitioners use some variant of this method (e.g. Samuelsson *et al.*, 2004). Defining the ridge width in terms of the mass of the given point models the variation of resolution along the m/z scale (Herzog *et al.*, 2011) (Fig. 3). The combined spectra (Fig. 2) are sliced into adjacent bins of width $\frac{m/z}{resolution}$, where m/z is the m/z at the current point and *resolution* is the resolution of the machine. Each bin is then treated as a ridge.

3.2 Modified bin agglomeration

Modified bin agglomeration (MBA) uses a series of decisions based on the shape of intensity histogram bins to partition the data into ridges. First, the data are binned according to the Fixed-Width algorithm, except with a user-defined bin width whose default is 5 ppm for the Orbitrap XL (Fig. 4). After this initial binning, the contiguous bins demarcated by empty bins are considered ridges. Note the difference between this and the Fixed-Width algorithm, which considers hard contiguous bin intervals as ridges irrespective of the content of each bin. At this point, if the user has selected the 0 option, the algorithm is complete.

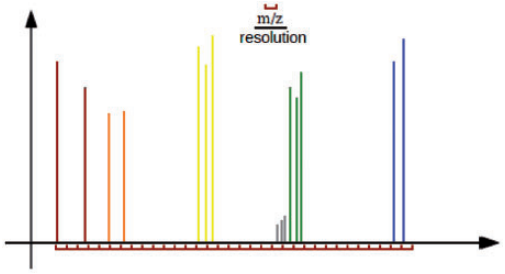

Fig. 3. Fixed-Width segmentation. The combined spectra (Fig. 2) are sliced into bins of width $\frac{m/z}{resolution}$. Note how fixed width has neither means of detecting data density or comparing the intensity of points. The shadow ridge (gray) is indistinguishable from the green ridge next to it, despite the intensity difference. Also, the hard bin limits segment-observed ridges that happen to fall on both sides of a bin interval. The colors delineate observed ridges. The red segments along the *x*-axis indicate bin boundaries. See online version for color

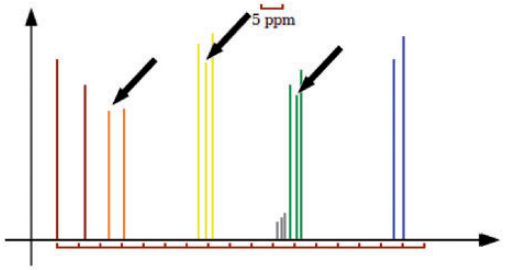

Fig. 4. Modified bin agglomeration segmentation. The combined spectra (Fig. 2) is sliced into bins of user-defined width (default 5 ppm). MBA then segments existing bins into disparate ridges at local minima (black arrows). The colors delineate observed ridges. See Figure 5 for more detail on MBA bin splitting. See online version for color

There are two other options available: share and greedy_y. Both options split all ridges where the sum of the intensities of each bin forms a local minima within a series of contigous bins. The difference between the share and greedy_y options consists of how these local minima are treated (Fig. 5).

3.3 Statistical agglomeration

Statistical agglomeration (SA) bases bin agglomeration decisions on statistical analysis of the data. The approach here is to treat ridges as distributions and bins of data as samples from those distributions. Although there is no guarantee that the samples being tested are normally distributed, we make this assumption to use the *t*-test. Ridges (distributions) whose means are not statistically different according to this test are combined iteratively until all remaining ridges are statistically different with high confidence.

As with the previous methods, the data are first sorted by ascending m/z and split into bins of size m/z_{window} (Fig. 6):

$$m/z_{window} = resolution \times 10^{-7} \tag{10}$$

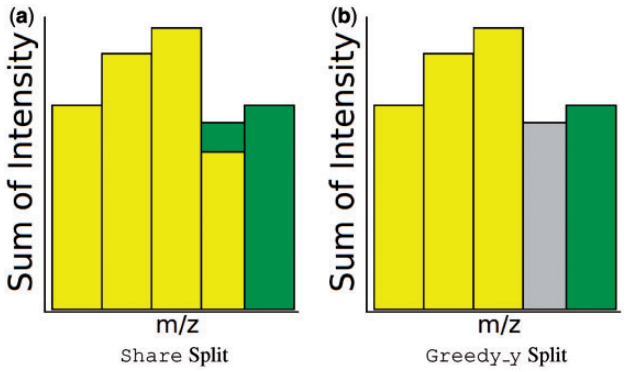

Fig. 5. MBA bin splitting. After segmenting all points into fixed interval bins and creating initial ridges of each contiguous segment bounded by empty bins, the MBA algorithm further divides ridges by considering local minima. With the share method (**a**), the local minimum is split among adjoining ridges proportional to the neighboring ridges' intensities. The greedy_y method (**b**) awards the entire disputed bin to the adjoining ridge of greatest total intensity. The bars in this figure represent histograms of the intensity of the points in the assigned bins, not the component points themselves. See online version for color

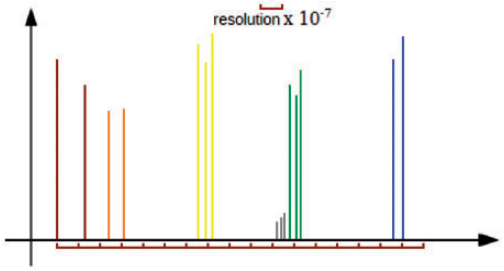

Fig. 6. Statistcal agglomeration segmentation. The combined spectra (Fig. 2) are sliced into bins of width *resolution* $\times 10^{-7}$. The colors delineate observed ridges. The red segments along the *x*-axis indicate bin boundaries. See online version for color

This formula was empirically derived from observation of several lipid samples to yield a good balance between minimal window size and sufficient size to estimate ridge statistics, and it should be applicable across many mass spectrometers.

After the initial bin assignment, starting at the lowest m/z value, adjacent bins are subjected to a Welch *t*-test (Welch, 1947) [we use the Welch *t*-test because the samples (bins) have potentially different sizes and variances] to test the hypothesis that the two sample distributions have the same mean:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{11}$$

where $\overline{X_i}$, s_i^2 and N_i are the *i*th sample mean, sample variance and sample size, respectively. The degrees of freedom are approximated using the Welch–Satterthwaite equation (Satterthwaite, 1946):

$$v = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{s_1^4}{N_1^2 \cdot (N_1 - 1)} + \frac{s_2^4}{N_2^2 \cdot (N_2 - 1)}} \tag{12}$$

For each potential bin agglomeration, the *P*-value is obtained from a *t*-distribution for a two-tailed test for the computed *t* and *v* values [Equations (11) and (12)] to validate the null hypothesis that the ridge means are equal. If the $P > 0.01$, meaning the confidence that they are different is <99%, we accept the null hypothesis and combine the bins being tested. To accommodate a test of both the m/z and intensity differences of the considered bins, each tested bin pair is subjected to two *t*-tests, one using the m/z data and one using the intensity data. As an overall measure of confidence, we use the maximum *P*-value for the two *t*-tests. The approach here is to be no more confident than our least confident *t*-test dimension (intensity or m/z). This design decision provides an implicit awareness of situations, which would be deceptive if the minimum *P*-value were used as an overall measure of confidence, such as when two bins have a very similar m/z values but different intensities. This situation, which we call *shadow ridges*, occurs surprisingly often when a low-intensity ridge appears directly adjacent to a high-intensity ridge. This approach also helps discriminate in cases when two bins that should not be combined are similar in average intensity. This is a common occurrence at low intensities. In this case, the lack of confidence in the m/z dimension will prevent combination of the two ridges.

In the event that the two bins under consideration are combined, the resulting agglomerated bin is considered as a single bin in the next iteration's comparison with the next bin in ascending m/z order. If they are not combined, the first bin in m/z order remains unchanged, and with the next iteration the second bin is compared with the next subsequent bin in ascending m/z order (Fig. 7). The entire algorithm runs in just one pass, resulting in O(*n*) performance, where *n* is the number of bins.

For post-processing noise removal, we use an established noise filtering method where all points with intensities below the estimated noise level (signal to noise ratio (s/n) = 1) are labeled as noise and removed. This method is borrowed from Samuelsson *et al*., but we modify the quantitation of noise from an intensity level to a frequency count, which is more robust to lower intensity signals (Samuelsson *et al*., 2004). This approach rests on the assumption that noise points are distributed uniformly, and thus should be equally distributed across the initial bins. The expected noise level is one noise point per bin.

3.4 Xcalibur

Xcalibur is a propriety mass spectrometry software platform from Thermo Scientific. Because Xcalibur will not accept data in the community standard mzML format, we were unable to use it on the Noyce synthetic dataset (Deutsch, 2008). However, the raw data of the Sample_3 datasets were analyzed using Xcalibur 2.1.

4 RESULTS

SA generally outperforms the other methods under consideration across all datasets on both the qualitative and quantitative measures considered in this study.

SA and MBA outperform all other methods on NTPD m/z (Fig. 8). MBA had a slightly lower NTPD rate on Sample_3_750-800, whereas SA outperformed all other methods

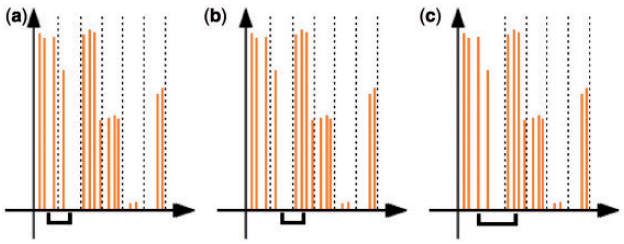

Fig. 7. SA bin agglomeration. After sorting the data by m/z value, and assigning data points to bins of fixed width, a *t*-test is conducted on the intensity and m/z means of the first two bins (**a**). If either of the *t*-tests fails to show a high confidence that the means are different, the bins are not combined and the algorithm considers the next two bins for agglomeration (**b**). Otherwise, the two bins are agglomerated, and the algorithm considers the agglomerated bin and the next bin for agglomeration (**c**). Dotted lines indicate ridge boundaries. See online version for color

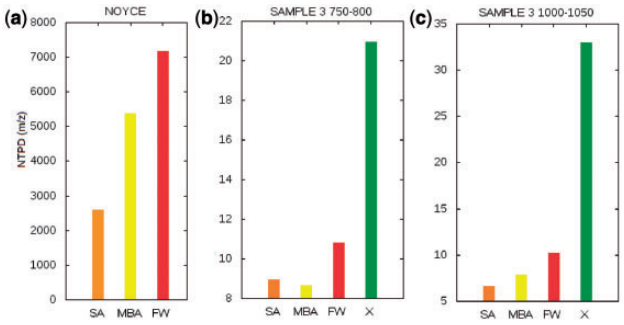

Fig. 8. NTPD—m/z. NTPD is a difference metric that compares the predicted ridge to the nearest observed ridge. Here, we compare the ridges' m/z values resulting from each of the four methods (Statistical Agglomeration, Modified Bin Agglomeration and Fixed Width) using the (**a**) Noyce, (**b**) Sample_3_750-800 and (**c**) Sample_3_1000-1050 datasets. On average, SA provides a 38% reduction in error from Fixed Width and provides a 68% improvement over Xcalibur for the two datasets for which Xcalibur's propriety data restrictions precluded measurement. Note the different scales. See online version for color

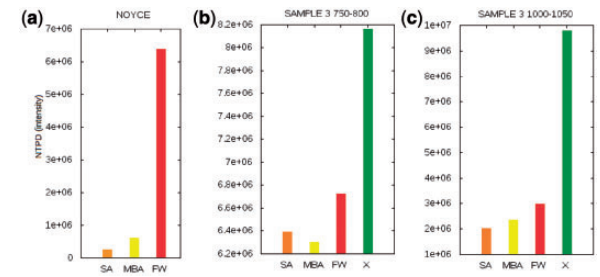

Fig. 9. NTPD—Intensity. Here we compare predicted ridge intensities to the nearest observed ridge for each of the four methods (Statistical Agglomeration, Modified Bin Agglomeration and Fixed Width) using the (**a**) Noyce, (**b**) Sample_3_750-800 and (**c**) Sample_3_1000-1050 datasets. SA outperforms the other methods on average, providing a 51% error reduction from Xcalibur for the two measurable datasets given Xcalibur's proprietary data restrictions. SA provides a 44% reduction on average over Fixed Width. Note the different scales. See online version for color

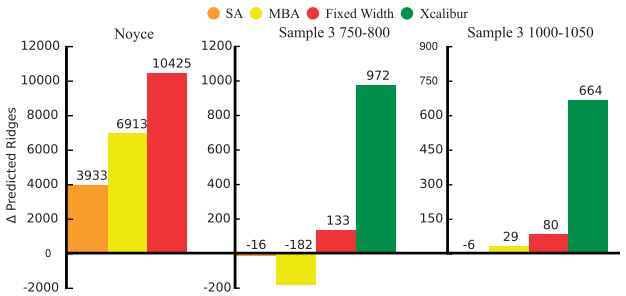

Fig. 10. Δ Number of ridges predicted by method. Each bar represents the difference from the actual number of ridges for each of the four methods (Statistical Agglomeration, Modified Bin Agglomeration and Fixed Width) summed across all datasets. SA's number of predicted ridges is much closer to the observed number than any other method. Xcalibur predicted far more ridges than any other method. Because Xcalibur only accepts data in its proprietary format, the results are not available for the Noyce dataset. Note the different scales. See online version for color. See online version for color

on the other two datasets. The relative performance was identical for NTPD intensity, with the exception being more disparity between the SA and MBA scores and fixed width on the Noyce dataset (Fig. 9a). Xcalbur's NTPD is dramatically higher for both NTPD intensity and NTPD m/z than all other methods on the two datasets that were comparable given Xcalibur's proprietary data limitations.

SA predicted the number of ridges far more accurately than any other method tested, including Xcalibur, which was furthest from the actual number of ridges (Fig. 10). MBA was second best on average at predicting the correct number of ridges.

On average, each of the three methods performs rather similarly on purity. The scores averaged across all three datasets are 0.73, 0.7 and 0.74 for SA, MBA and Fixed Width respectively (Supplementary Material). Because we are ignoring all noise points (real or assigned), and because Fixed Width produces the narrowest ridges, it is not surprising that Fixed Width performed so well on purity.

The NMI scores averaged across all three datasets are 0.95, 0.96 and 0.93 for SA, MBA and Fixed Width, respectively (Supplementary Material). It is surprising that they are so close, but this is likely a result of the modifications to this metric to handle noise.

Each of the three methods performs inconsistently on SSE. SA outperforms the other methods on both Sample 3 datasets for m/z SSE, but MBA has a dramatically lower SSE for the Noyce dataset than either of the other methods (Fig. 11). Fixed Width has a dramatically lower intensity SSE than either of the other methods on the Noyce dataset, but only slightly less SSE than SA on the Sample_3_750-800 dataset (Supplementary Material). MBA noticeably outperforms other methods on the Sample_3_1000-1050 dataset.

Although the aforementioned metrics should give an overall quantitative measure of the performance of each method, the segments of the spectra in Figures 12 and 13 provide a qualitative

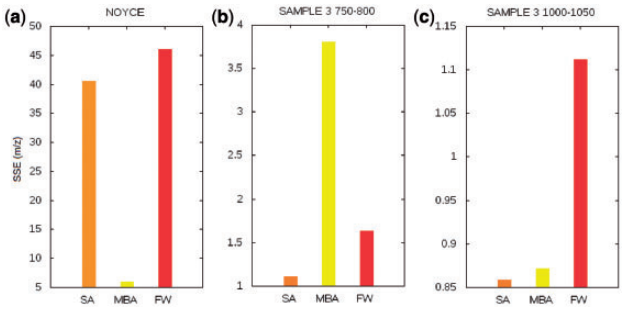

Fig. 11. Sum-squared error (SSE)—m/z. The SSE of each of the three of the four methods (Statistical Agglomeration, Modified Bin Agglomeration and Fixed Width) is measured for each of the three datasets (**a**) Noyce, (**b**) Sample_3_750-800 and (**c**) Sample_3_1000-1050 datasets. SA outperforms the other methods on Sample_3_750-800 and Sample_3_1000-1050, but MBA outperforms the other methods on the Noyce dataset. SA's average error is 23% lower than Fixed Width. This metric could not be measured for Xcalibur's ridge assignments. Note the different scales. See online version for color

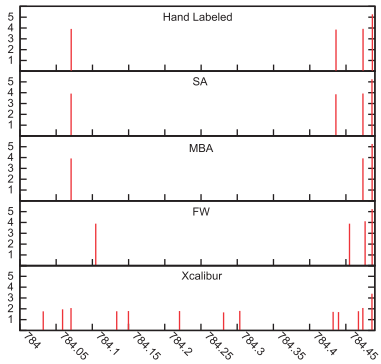

Fig. 12. Peak summarization of sample 3: 784-785. Note: all intensities have been log-transformed for fit. See online version for color

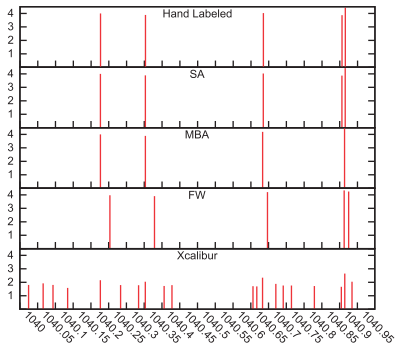

Fig. 13. Peak summarization of sample 3: 1040-1041. Note: all intensities have been log-transformed for fit. See online version for color

assessment of each method on the Sample 3 datasets (Supplementary Material for quantitative results for the Noyce dataset). The pattern that emerges across datasets is that, at least on these random segments, SA consistently summarizes ridges exactly or very close to the hand annotation. MBA also performs well. Fixed Width is not consistent in performance but usually adds extra ridges and/or shifts m/z values of ridges substantially. Across both Sample 3 datasets, Xcalibur drastically increases the number of ridges in the segment. Xcalibur's predicted ridges are also notably less intense than the hand-annotated dataset.

The Supplementary information provides more metric results and an in-depth discussion of how noise is treated in this article.

5 DISCUSSION

Fixed Width, to our knowledge the only extant algorithmic solution to this problem, is simple to code, yet has some obvious limitations. In mass spectrometry, the intra-sample resolution is inherently variable (Schwudke *et al.*, 2011). At least for the Orbitrap, low-intensity signal groups are more dispersed, whereas high-intensity signal groups have less m/z variance. Any fixed-width solution will either chop low-intensity ridges into incorrect component ridges, incorrectly agglomerate high-intensity ridges or both. As shown in the results, fixed-width methods significantly overestimate the number of ridges, cascading error downstream into identification and quantitation.

MBA attempts to provide robust means for dealing with ridges that overlap, and builds on the idea of Fixed Width binning by agglomerating any adjacent non-empty bins. Although the initial fixed width and the choice of which bin splitting options to use are parameters that must be determined and set by the operator, the information in manufacturer specifications, such as resolution, should assist in deciding the MBA parameters. In practice, the machine calibration to which the specifications are tied is not always the setup desired for the practitioner due to time requirements, desire to use MS/MS and so forth. Also, the true machine resolution can vary widely outside of the m/z value the specification is provided for. However, practical experience may assist in knowing when the manufacturer specs are sufficient and what changes need to be made when they are not.

Because each ridge can be a different width, SA addresses the problem of bin size in a flexible data-driven manner. The ridge agglomeration procedure is statistically driven using the data itself, handling problems like overlapping ridges and avoiding the need for users to set parameters or for *a priori* knowledge about the dataset. Noise filtering allows for the avoidance of boundary conditions found in fixed-width methods such as ridges with just one data point. We consider $s/n = 1$ to be a useful *a priori* setting, as it was the ideal setting across all three of our datasets. SA's ability to predict a far more accurate number of ridges than the other methods suggests it will increase accuracy in downstream processes over the current methods used, including Xcalibur (Fig. 10).

One troubling observation from this study is the difficulty in accurately assessing intensity of discovered ridges. Both species identification and quantitation require an accurate intensity measurement. Yet, even SA's performance is simply the best of several inaccurate methods. Given the amount of lipid quantitation performed currently, and also the state-of-the-art, better methods of estimating intensity are needed.

We have described the need for accurate ridge summarization in direct injection lipidomics samples. Interestingly, despite the importance of accuracy in this first step of the analysis pipeline, there has been no study of solutions to this version of the ridge

summarization problem to our knowledge. We present our estimate of what is currently done in the community and also propose two novel algorithms, MBA and SA, for resolving ridges in shotgun lipidomics samples. We show that SA outperforms open source and proprietary methods on average in a measure of ridgewise error, NTPD, on three datasets. We also show that SA significantly outperforms the proprietary program Xcalibur on the two datasets for which we could use Xcalibur.

Incorporation of SA into existing analysis pipelines could drastically improve downstream quantitation and identification results in a variety of lipidomics experiments. Future work should continue improving our capacity to produce summarized ridges that more accurately estimate intensity. In light of the recent calls for greater reproducibility in mass spectrometry (Wilkins *et al*., 2006), and to foster development of improved algorithms, these datasets and the SA algorithm (with ample documentation) are available freely for non-commercial use at http://github.com/optimusmoose/statistical_agglomeration.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Brendan Coutu for his assistance in the annotation of the datasets and Ryan Williamson for his assistance with some of the visualizations used in this article.

Funding: This work was supported by the National Science Foundation Graduate Research Fellowship [DGE-0750759] and institutional funds.

Conflict of Interest: none declared.

REFERENCES

Deutsch,E. (2008) mzML: a single, unifying data format for mass spectrometer output. *Proteomics*, **8**, 2776–2777.

Ejsing,C.S. *et al*. (2006) Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning. *Anal. Chem*., **78**, 6202–6214.

Ekroos,K. *et al*. (2002) Quantitative profiling of phospholipids by multiple precursor ion scanning on a hybrid quadrupole time-of-flight mass spectrometer. *Anal. Chem*., **74**, 941–949.

Frank,A.M. *et al*. (2008) Clustering millions of tandem mass spectra. *J. Proteome Res*., **7**, 113–122.

Han,X. and Gross,R.W. (2005) Shotgun lipidomics: electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples. *Mass Spectrom. Rev*., **24**, 367–412.

Herzog,R. *et al*. (2011) A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language. *Genome Biol*., **12**, R8.

Liu,J. *et al*. (2007) Methods for peptide identification by spectral comparison. *Proteome Sci*., **5**, 3.

Noyce,A.B. *et al*. (2013) Mspire-simulator: LC-MS shotgun proteomic simulator for creating realistic gold standard data. *J. Proteome Res*, epub ahead of print.

Orešič,M. (2009) Bioinformatics and computational approaches applicable to lipidomics. *Eur. J. Lipid Sci. Technol*., **111**, 99–106.

Samuelsson,J. *et al*. (2004) Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics*, **20**, 3628–3635.

Satterthwaite,F.E. (1946) An approximate distribution of estimates of variance components. *Biometrics*, **2**, 110–114.

Schwudke,D. *et al*. (2006) Lipid profiling by multiple precursor and neutral loss scanning driven by the data-dependent acquisition. *Anal. Chem*., **78**, 585–595.

Schwudke,D. *et al*. (2011) Shotgun lipidomics on high resolution mass spectrometers. *Cold Spring Harb. Perspect. Biol*., **3**, a004614.

Song,H. *et al*. (2007) Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching. *J. Am. Soc. Mass Spectrom*., **18**, 1848–1858.

Sturm,M. and Kohlbacher,O. (2009) Toppview: An open-source viewer for mass spectrometry data. *J. Proteome Res*., **8**, 3760–3763.

Watson,A.D. (2006) Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Lipidomics: a global approach to lipid analysis in biological systems. *J. Lipid Res*., **47**, 2101–2111.

Welch,B.L. (1947) The generalization of Student's problem when several different population varlances are involved. *Biometrika*, **34**, 28–35.

Wilkins,M.R. *et al*. (2006) Guidelines for the next 10 years of proteomics. *Proteomics*, **6**, 4–8.