BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–7

Statistical Agglomeration: Peak Summarization for Direct Infusion Lipidomics

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

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

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 isolation 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% (m/z) and 44% (intensity) compared to the ad-hoc method over 3 data sets. Pointwise error is reduced by 23% (m/z). Compared to Xcalibur, our statistical agglomeration algorithm produces 68% less m/z error and 51% less intensity error on average on two comparable data sets.

Availability: The source code for Statistical Agglomeration is 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, jtprince@chem.byu.edu

1 INTRODUCTION

Direct infusion (injection) lipidomics, sometimes called "shotgun" lipidomics for it's similarity to shotgun genomics, is an emerging but well studied field (Watson, 2006; Ekroos *et al.*, 2002; Ejsing *et al.*, 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). Since 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 two-dimensional vector of unique peaks 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's more, technical and mechanical limitations in the mass spectrometer inculcate even more error into the output. Accurately estimating the true peak values from the resulting output file is a nontrivial challenge (see Figure 1).

In order to identify and quantify each lipid, it's component peaks must somehow be isolated one from another, and the additive noise peaks removed. We will call this process *peak summarization*. Only after peak summarization can the isotopic envelopes be compared with theoretical databases in order to identify and quantify the individual lipids in the sample.

The necessity of a solution for the peak 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 peak summarization be found in every shotgun lipidomics study (Samuelsson *et al.*, 2004). However, it is frequently left unmentioned (e.g. (Orešič, 2009), (Song *et al.*, 2007), and (Ejsing *et al.*, 2006)). Although direct infusion methods have been around since the mid-1990s, we are only aware

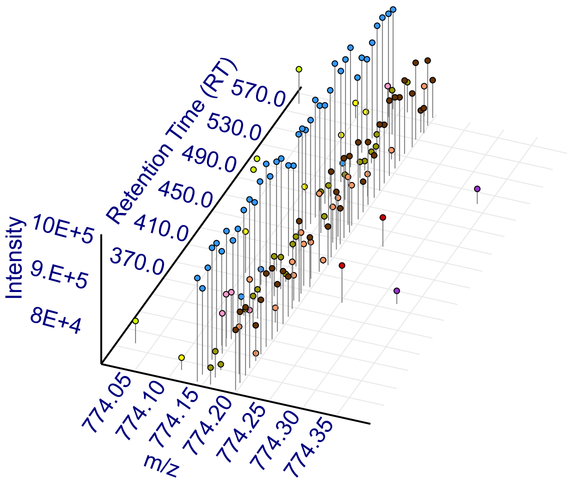


Fig. 1: A typical direct infusion lipidomics sample. The lack of consistent repetition in data points in the RT dimension and the abundance of noise in each of the three dimensions make accurate peak summarization difficult. The colors delineate true peaks.

[*] to whom correspondence should be addressed

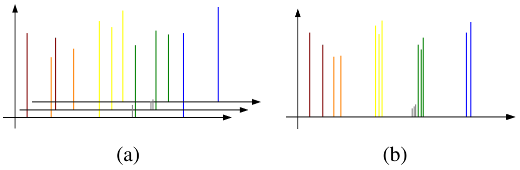


Fig. 2: Scan Combination. Here (a) multiple scans are combined into (b) one list of (m/z, intensity) pairs by removing the retention time (RT) dimension. The colors delineate peaks.

of two published solutions to this segment of the quantitation pipeline. The first is that of treating a survey scan as a true peak 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 peaks would be less than ideal, as the scan would include many peaks with incorrect m/z and intensity and exclude many other true peaks (see Figure 1). The second, a more robust approach, applies to shotgun lipidomics a technique that has been used in several proteomics studies (Liu *et al.*, 2007; Frank *et al.*, 2008). This approach, which we label the fixed width algorithm, averages scans across the retention time 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 peak summarization problem and evaluate them against both synthetic and real-world peak summarization problems. We also provide the first comparative performance analysis of Xcalibur and the fixed width algorithm on the peak summarization problem.

2 SYSTEM AND METHODS

We use a representative sample of three labeled data sets 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 paper were evaluated on one synthetic data set and two real world, hand labeled data sets.

The Noyce data set is a synthetic data set constructed as described in (Noyce *et al.*, view) with sampling rate 1, noise density factor 500, and one dimension mode.

Sample_3_750-800 and Sample_3_1000-1050 are two m/z intervals of a rat soleus lipid extract (see supplement for experimental protocol). Each peak in these data sets were 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 data sets used in lipidomics can be represented as a list of points where each point is an (m/z, intensity, RT) triplet. For the purposes of the algorithms detailed here, the RT dimension is ignored, reducing the problem to two dimensions (see Figure 2).

2.2 Metrics

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

In what follows, we define $\mathbf{\Omega}$ as the set of predicted peaks, $\mathbf{C}$ as the set of true peaks, $\mathbf{I}$ as the set of data points, ω_k as the set of points in the kth predicted peak in $\mathbf{\Omega}$, and c_j as the set of points in the jth true peak in $\mathbf{C}$. We define the intensity, ω_{int}, of a predicted peak ω as the total of the intensities of the peak's assigned points:

$$\omega_{int} = \sum_{i \in \omega} i_{int} \tag{1}$$

and the m/z value, $\omega_{m/z}$, of ω as weighted mean m/z values of the peak's assigned points is defined as:

$$\omega_{m/z} = \sum_{i \in \omega} i_{m/z} \frac{i_{int}}{\omega_{int}} \tag{2}$$

Let $\omega^i_{m/z}$ denote the m/z of the predicted peak containing point i and $c^i_{m/z}$ the m/z of the true peak containing point i.

NORMALIZED TRUE PEAK DISTANCE (NTPD). NTPD is a metric we developed for this task which indicates the normalized m/z or intensity difference between the predicted peaks and the nearest true peaks. For m/z NTPD the equation is,

$$\text{NTPD}(\mathbf{\Omega}, \mathbf{C}) = \frac{1}{\min(|\mathbf{\Omega}|, |\mathbf{C}|)} \sum_{k \in \mathbf{\Omega}} \min_{j \in \mathbf{C}} (|\omega^k_{m/z} - c^j_{m/z}|) \tag{3}$$

Intensity NTPD is calculated using the same equation with $\omega^k_{m/z}$ and $c^j_{m/z}$ replaced with ω^k_{int} and c^j_{int}.

The normalizing term controls score inflation whether the error is in predicting too many or too few peaks. The significance of this metric is reflected in its clinical relevancy. This per-peak 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 PEAKS. In downstream algorithms, each estimated peak 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 peaks versus the number of actual peaks.

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

$$\text{purity}(\mathbf{\Omega}, \mathbf{C}) = \frac{1}{|\mathbf{I}|} \sum_{k \in \mathbf{\Omega}} \max_{j \in \mathbf{C}} |\omega_k \cap c_j| \tag{4}$$

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

NORMALIZED MUTUAL INFORMATION (NMI). NMI allows the quantitation of the trade off between number of predicted peaks and the quality of predicted peaks.

$$\mathrm{NMI}(\mathbf{\Omega}, \mathbf{C}) = \frac{\mathrm{I}(\mathbf{\Omega}, \mathbf{C})}{[\mathrm{H}(\mathbf{\Omega}) + \mathrm{H}(\mathbf{C})]/2} \tag{5}$$

where I is mutual information, given by

$$\mathrm{I}(\mathbf{\Omega}, \mathbf{C}) = \sum_{k\in\mathbf{\Omega}} \sum_{j\in\mathbf{C}} \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \tag{6}$$

and H is entropy, given by

$$\mathrm{H}(\mathbf{\Omega}) = -\sum_{k\in\mathbf{\Omega}} \frac{|\omega_k|}{|\mathbf{I}|} \log \frac{|\omega_k|}{|\mathbf{I}|} \tag{7}$$

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

SUM SQUARED ERROR (SSE). 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:

$$\mathrm{SSE}(\mathbf{\Omega}, \mathbf{C}) = \sum_{i\in I} (\omega^i_{m/z} - c^i_{m/z})^2 \tag{8}$$

Intensity SSE is calculated in the same fashion, with intensity replacing m/z in Equation 8.

These metrics had to be modified to deal with the notion of noise points, which are not inherent in clustering problems (discussed in supplemental information).

3 ALGORITHMS

While both methods proposed as well as the fixed width method follow the peak summarization paradigm by combining multiple scans (see Figure 2), each of the three methods diverges in the way the peaks are segmented once combined into one spectra.

3.1 Fixed Peak Width Method

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

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 peaks. First, the data is binned according to the Fixed Width algorithm, except with a user-defined bin width whose default is 5ppm for the Orbitrap XL (see Figure 4). After this initial binning, the contiguous bins demarcated by empty bins are considered peaks. Note the difference between this and the Fixed Width algorithm, which considers hard contiguous bin intervals as peaks irrespective of the content of each bin. At this point, if the user has selected the `zero` option, the algorithm is complete.

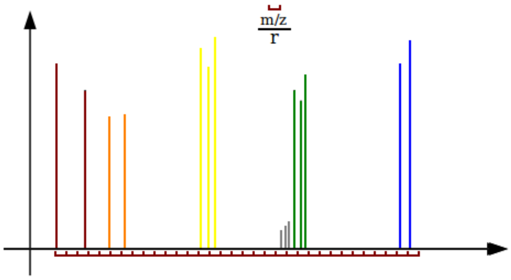


Fig. 3: Fixed Width Segmentation. The combined spectra (see Figure 2) are sliced into bins of width $\frac{m/z}{r}$, where r is the resolution of the instrument. Note how fixed width has no means of detecting data density, nor comparing the intensity of points. The shadow peak (gray) is indistinguishable from the geen peak next to it, despite the intensity difference. Also, note how the hard bin limits segment true peaks that happen to fall on both sides of a bin interval. The colors delineate true peaks. The red segments along the x-axis indicate bin boundaries.

There are two other options available: `share` and `greedy_y`. Both options split all peaks where the sum of the intensities of each bin form 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 (see Figure 5).

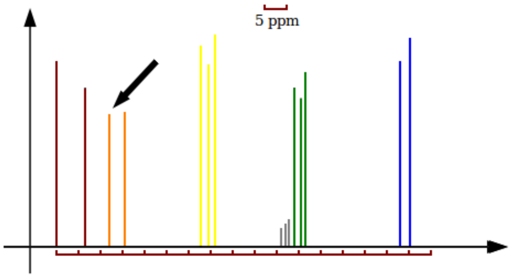
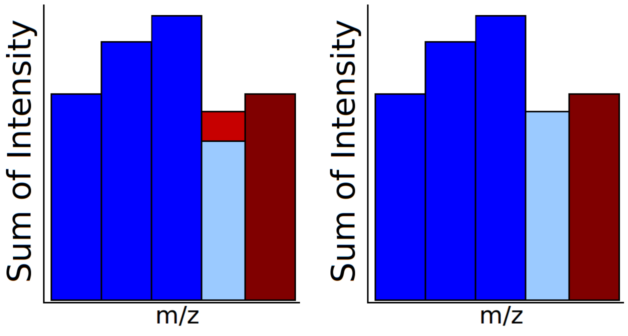


Fig. 4: Modified Bin Agglomeration Segmentation. The combined spectra (see Figure 2) is sliced into bins of user-defined width (default 5ppm). MBA then segments existing bins into disparate peaks at local minima (black arrow). The colors delineate true peaks. See Figure 5 for more detail on MBA bin splitting.

3.3 Statistical Agglomeration

Statistical Agglomeration (SA) bases bin agglomeration decisions on statistical analysis of the data. The approach here is to treat peaks as distributions and bins of data as samples from those distributions. Although there is no guarantee that the samples being tested are



(a) `Share` Split (b) `Greedy_y` Split

Fig. 5: MBA Bin Splitting. After segmenting all points into fixed interval bins and creating initial peaks of each contiguous segment bounded by empty bins, the MBA algorithm further divides peaks by considering local minima. With the `share` method (a), the local minimum is split among adjoining peaks proportional to the neighboring peaks' intensities. The `greedy_y` method (b) awards the entire disputed bin to the adjoining peak of greatest total intensity. Note that the bars in this figure represent histograms of the intensity of the points in the assigned bins, not the component points themselves.

normally distributed, we make this assumption in order to use the t-test. Peaks (distributions) whose means are not statistically different according to this test are combined iteratively until all remaining peaks are statistically different with high confidence.

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

$$m/z_{window} = \text{resolution} \times 10^{-7} \tag{9}$$

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 peak 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:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{10}$$

where $\bar{X}_i$, 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):

$$v = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{s_1^4}{N_1^2 \cdot (N_1 - 1)} + \frac{s_2^4}{N_2^2 \cdot (N_2 - 1)}} \tag{11}$$

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 (see Eq. 10, 11) to validate the null hypothesis that the peak means are equal. If the p value is greater than 0.01, meaning the confidence that they are different is less than 99%, we accept the null hypothesis and combine the bins being tested. Note that, in order 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 very different intensities. This situation, which we call *shadow peaks*, occurs surprisingly often when a low intensity peak appears directly adjacent to a very high intensity peak. 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 peaks.

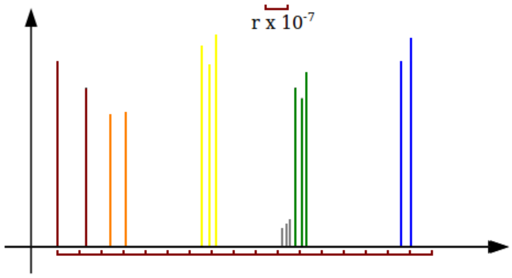


Fig. 6: Statistcal Agglomeration Segmentation. The combined spectra (see Figure 2) is sliced into bins of width $r \times 10^{-7}$, where r is the resolution of the instrument. The colors delineate true peaks. The red segments along the x-axis indicate bin boundaries.

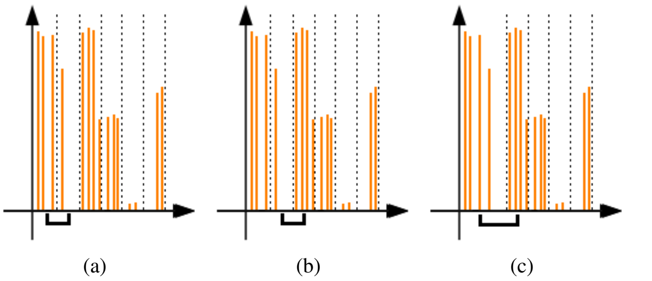


Fig. 7: SA Bin Agglomeration. After sorting the data by m/z value, and assigning data points to bins of fixed width, a t-test is conducted on the intensity and m/z means of the first two bins (a). If either of the t-tests fail to show a high confidence that the means are different, the bins are not combined 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 peak boundaries.

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 to 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 (see Figure 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. Since Xcalibur will not accept data in the community standard mzML format, we were unable to use it on the Noyce synthetic data set (Deutsch, 2008). However, the raw data of the Sample_3 data sets were analyzed using Xcalibur 2.1.

4 RESULTS

SA and MBA outperform all other methods on NTPD m/z (see Figure 8). MBA had a slightly lower NTPD rate on Sample_3_750-800, while SA outperformed all other methods on the other two data sets. 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 data set (see Figure 9(a)). Note that Xcalbur's NTPD is dramatically higher for both NTPD intensity and NTPD m/z than all other methods on the two data sets that were comparable given Xcalibur's proprietary data limitations.

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

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

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

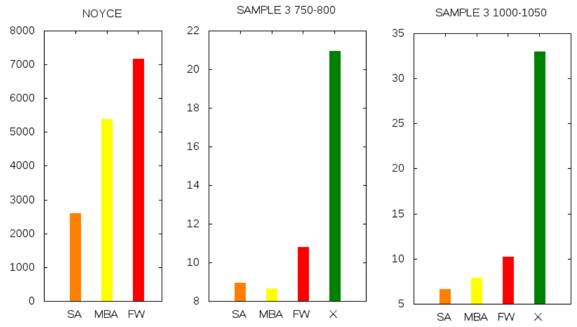
Each of the three methods performs inconsistently on SSE. SA outperforms the other methods on both Sample 3 datasets for m/z SSE, but MBA has a dramatically lower SSE for the Noyce dataset then either of the other methods (see Figure 11). Fixed Width has a dramatically lower intensity SSE than either of the other methods on the Noyce data set, but only slightly less SSE than SA on the Sample_3_750-800 data set (see supplemental material). MBA noticably outperforms other methods on the Sample_3_1000-1050 data set.

While the above reported metrics should give an overall quantitative measure of the performance of each method, the segments of the spectra in Figures 12, 13, and 14 provide a qualitative assessment of each method. The pattern that emerges across data sets is that, at least on these random segments, SA consistly summarizes peaks exactly or very close to the hand annotation. MBA also performs well. Fixed Width is not consistent in performance but usually adds extra peaks and/or shifts m/z values of peaks substantially. Across both Sample 3 datasets, Xcalibur drastically increases the number of peaks in the segment. Xcalibur's predicted peaks are also notably less intense than the hand annotated data set.

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 while high intensity signal groups have less m/z variance. Any fixed width solution will either chop low intensity peaks into incorrect component peaks, incorrectly agglomerate high intensity peaks, or both. As shown in the results, fixed width methods significantly overestimate the number of peaks, cascading error downstream into identification and quantitation.

MBA attempts to provide robust means for dealing with peaks 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



(a) (b) (c)

Fig. 8: Normalized True Peak Distance (NTPD) - m/z. NTPD is a difference metric that compares the predicted peak to the nearest true peak. Here we compare the peaks' m/z values. On average, SA provides a 38% reduction in error from Fixed Width and provides a 68% improvement over Xcalibur for the two data sets for which Xcalibur's propriety data restrictions did not preclude measurement. Note the different scales.

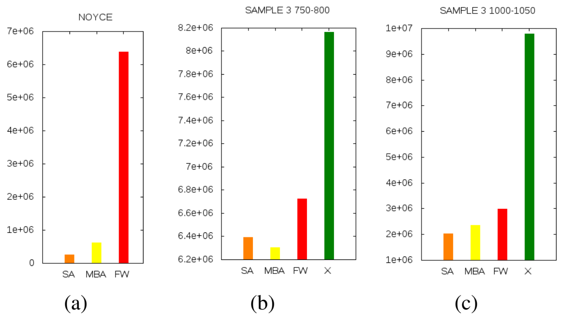


Fig. 9: Normalized True Peak Distance (NTPD) - Intensity. Here we compare predicted peak intensities to the nearest true peak. SA outperforms the other methods on average, providing a 51% error reduction from Xcalibur for the two measurable data sets given Xcalibur's proprietary data restrictions. SA provides a 44% reduction on average over Fixed Width. Note the different scales.

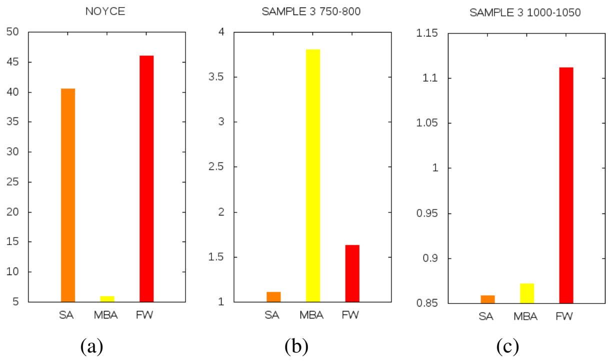


Fig. 11: Sum Squared Error (SSE) - m/z. SA outperforms the other methods on Sample_3_750-800 and Sample_3_1000-1050, but MBA outperforms the other methods on the Noyce data set. SA's average error is 23% lower than Fixed Width. This metric could not be measured for Xcalibur's peak assignments. Note the different scales.

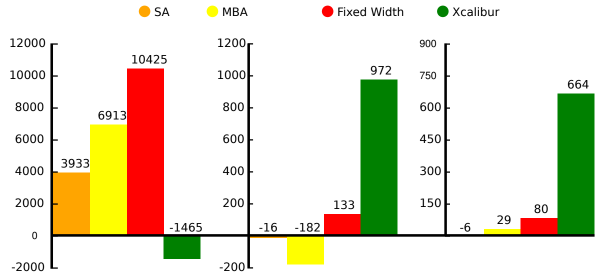


Fig. 10: Δ Number of Peaks Predicted by Method. Each bar represents the difference from the actual number of peaks. SA's number of predicted peaks is much closer to the true number than any other method. Xcalibur predicted far more peaks than any other method. Because Xcalibur only accepts data in its proprietary format, the results are not available for the Noyce data set. Note the different scales.

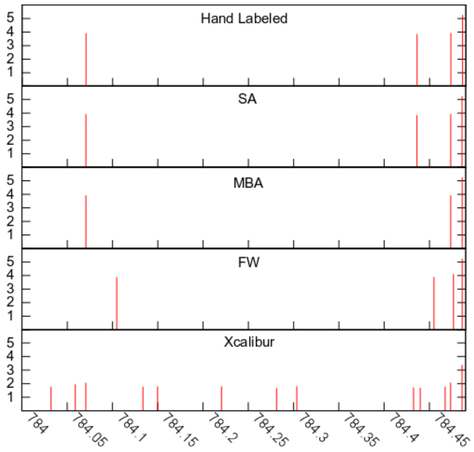


Fig. 12: Peak Summarization of Sample 3: 784-785. Note: all intensities have been log-transformed for fit.

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, etc. 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.

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

One troubling observation from this study is the difficulty in accurately assessing intensity of discovered peaks. 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 peak 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 peak summarization problem to our knowledge. We present our estimate of what is currently done in the community, and also propose two

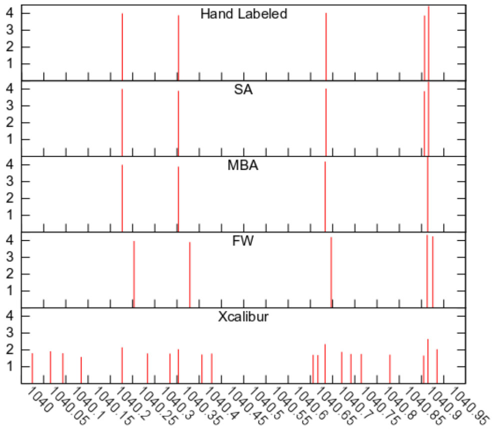


Fig. 13: Peak Summarization of Sample 3: 1040-1041. Note: all intensities have been log-transformed for fit.

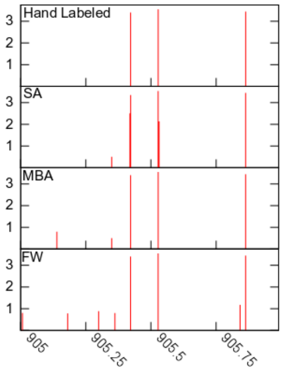


Fig. 14: Peak Summarization of the Noyce Data Set. Note: all intensities have been log-transformed for fit. Xcalibur could not be compared due to proprietary data restrictions.

novel algorithms, MBA and SA, for resolving peaks in shotgun lipidomics samples. We show that SA outperforms open source and proprietary methods on average in a measure of peakwise error, NTPD, on three data sets. We also show that SA significantly outperforms the proprietary program Xcalibur on the two data sets 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 peaks 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 data sets and the SA algorithm (with ample documentation) are available freely for non-commercial use at `http://github.com/optimusmoose/statistical_agglomeration`.

ACKNOWLEDGEMENT

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

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

6 REFERENCES

REFERENCES

Deutsch, E. (2008). mzML: A single, unifying data format for mass spectrometer output. *PROTEOMICS*, **8**(14), 2776–2777.

Ejsing, C. S., Duchoslav, E., Sampaio, J., Simons, K., Bonner, R., Thiele, C., Ekroos, K., and Shevchenko, A. (2006). Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning. *Analytical Chemistry*, **78**(17), 6202–6214.

Ekroos, K., Chernushevich, I. V., Simons, K., and Shevchenko, A. (2002). Quantitative profiling of phospholipids by multiple precursor ion scanning on a hybrid quadrupole time-of-flight mass spectrometer. *Analytical Chemistry*, **74**(5), 941–949.

Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A. (2008). Clustering millions of tandem mass spectra. *Journal of Proteome Research*, **7**(1), 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 Spectrometry Reviews*, **24**(3), 367–412.

Herzog, R., Schwudke, D., Schuhmann, K., Sampaio, J., Bornstein, S., Schroeder, M., and Shevchenko, A. (2011). A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language. *Genome Biology*, **12**, 1–25. 10.1186/gb-2011-12-1-r8.

Liu, J., Bell, A., Bergeron, J., Yanofsky, C., Carrillo, B., Beaudrie, C., and Kearney, R. (2007). Methods for peptide identification by spectral comparison. *Proteome Science*, **5**, 1–12. 10.1186/1477-5956-5-3.

Noyce, A. B., Dalgliesh, J., Taylor, R. M., Erb, K., Okuda, N., and Prince, J. T. ((Under Review)). Mspire-simulator: LC-MS shotgun proteomic simulator for creating realistic gold standard data. *Journal of Proteome Research*.

Orešič, M. (2009). Bioinformatics and computational approaches applicable to lipidomics. *European Journal of Lipid Science and Technology*, **111**(1), 99–106.

Samuelsson, J., Dalevi, D., Levander, F., and Rögnvaldsson, T. (2004). Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics*, **20**(18), 3628–3635.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**(6), pp. 110–114.

Schwudke, D., Oegema, J., Burton, L., Entchev, E., Hannich, J. T., Ejsing, C. S., Kurzchalia, T., and Shevchenko, A. (2006). Lipid profiling by multiple precursor and neutral loss scanning driven by the data-dependent acquisition. *Analytical Chemistry*, **78**(2), 585–595. PMID: 16408944.

Schwudke, D., Schuhmann, K., Herzog, R., Bornstein, S. R., and Shevchenko, A. (2011). Shotgun lipidomics on high resolution mass spectrometers. *Cold Spring Harbor Perspectives in Biology*, **3**.

Song, H., Hsu, F.-F., Ladenson, J., and Turk, J. (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. *Journal of the American Society for Mass Spectrometry*, **18**(10), 1848–1858.

Sturm, M. and Kohlbacher, O. (2009). Toppview: An open-source viewer for mass spectrometry data. *Journal of Proteome Research*, **8**(7), 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. *Journal of Lipid Research*, **47**(10), 2101–2111.

Welch, B. L. (1947). The generalization of Student's problem when several different population varlances are involved. *Biometrika*, **34**(1-2), 28–35.

Wilkins, M. R., Appel, R. D., Van Eyk, J. E., Chung, M. C. M., Görg, A., Hecker, M., Huber, L. A., Langen, H., Link, A. J., Paik, Y.-K., Patterson, S. D., Pennington, S. R., Rabilloud, T., Simpson, R. J., Weiss, W., and Dunn, M. J. (2006). Guidelines for the next 10 years of proteomics. *PROTEOMICS*, **6**(1), 4–8.