

Data and text mining

Warpgroup: increased precision of metabolomic data processing by consensus integration bound analysis

Nathaniel G. Mahieu^{1,2,*}, Jonathan L. Spalding^{1,3} and Gary J. Patti^{1,2,*}

¹Department of Chemistry, Washington University, St Louis, MO 63130, USA, ²Department of Medicine and

³Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on June 17, 2015; revised on September 4, 2015; accepted on September 22, 2015

Abstract

Motivation: Current informatic techniques for processing raw chromatography/mass spectrometry data break down under several common, non-ideal conditions. Importantly, hydrophilic liquid interaction chromatography (a key separation technology for metabolomics) produces data which are especially challenging to process. We identify three critical points of failure in current informatic workflows: compound specific drift, integration region variance, and naive missing value imputation. We implement the Warpgroup algorithm to address these challenges.

Results: Warpgroup adds peak subregion detection, consensus integration bound detection, and intelligent missing value imputation steps to the conventional informatic workflow. When compared with the conventional workflow, Warpgroup made major improvements to the processed data. The coefficient of variation for peaks detected in replicate injections of a complex *Escherichia Coli* extract were halved (a reduction of 19%). Integration regions across samples were much more robust. Additionally, many signals lost by the conventional workflow were ‘rescued’ by the Warpgroup refinement, thereby resulting in greater analyte coverage in the processed data.

Availability and implementation: Warpgroup is an open source R package available on GitHub at github.com/nathaniel-mahieu/warpgroup. The package includes example data and XCMS compatibility wrappers for ease of use.

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

Contact: nathaniel.mahieu@wustl.edu or gjpattij@wustl.edu

1 Introduction

Omics-scale separation/mass spectrometry approaches (e.g. LC/MS, GC/MS, CE/MS, etc.) generate large, 3D data sets consisting of elution time (rt), mass-to-charge ratio (m/z), and signal intensity information (Crutchfield *et al.*, 2010). Analytes are separated by their chemical characteristics prior to being introduced into the mass spectrometer (yielding rt). The mass spectrometer acts as a second dimension of separation and a detector, providing information on the accurate mass (m/z) and amount of each analyte (signal intensity). Each sample run can generate gigabytes of data representing tens of thousands of distinct analytes (Käll and Vitek, 2011).

The processing of raw data is a significant challenge and the conventional workflow consists of several steps. These steps include mass trace detection, chromatographic feature detection, inter-sample retention time drift correction, inter-sample grouping of common features (correspondence determination), and statistical analysis of feature groups (Patti *et al.*, 2012). A feature in this context refers to signal which displays a peak shape in both m/z and rt domains. The result of this data processing is quantification of all unique analytes detected across multiple sample runs.

Historically, most chromatography/mass spectrometry experiments have been performed with reversed-phase chromatography.

This well-established separation technique commonly generates Gaussian peak shapes and exhibits highly reproducible retention times. A simple retention mechanism based primarily on compound polarity also minimizes compound specific drift (Kele and Guiochon, 2000). One drawback to reversed-phase separation is a lack of retention for the highly polar compounds such as sugars and organic acids commonly of interest in metabolomic studies. As a result, many new separation chemistries have emerged under the umbrella term hydrophilic interaction liquid chromatography (HILIC), which aim to achieve separation of polar molecules (Buszewski and Noga, 2012). Unfortunately, analytes measured by HILIC separation exhibit a wide range of non-Gaussian peak shapes as well as larger, compound-specific retention time drift (Fuhrer and Zamboni, 2015). Current informatic approaches for metabolomics were primarily developed by using reversed-phase C18 chromatography, and even today most new advances are benchmarked solely on reversed-phase datasets (Tautenhahn *et al.*, 2008). Thus, the performance of these algorithms degrades when applied to HILIC datasets.

Detection of features and selection of integration regions is an initial and critical step of the informatic workflow (Cappadona *et al.*, 2012). In cases where peak shapes are simple and peaks exhibit large signal-to-noise ratios, the detection and integration of peaks is reproducible. Complex metabolomic datasets, however, contain a high proportion of poorly resolved and low-abundance peaks (Nikolskiy *et al.*, 2013). Additionally, the non-Gaussian peak shapes exhibited by a large portion of HILIC features impede the robust selection of integration bounds. These factors complicate peak detection and result in undetected features as well as integration bounds which describe different regions of a peak in each sample (Fig. 1A and B).

The second major informatic step is determination of correspondence. Current feature grouping techniques rely on the reproducible elution of compounds across multiple experimental runs. The elution time of m/z -rt pairs (i.e. features) is the key information used to associate the same compound detected in different runs (Vandenbogaert *et al.*, 2008). In practice, elution times vary from sample to sample due to many factors (Quarry *et al.*, 1984). This necessitates correction of retention time drift prior to grouping. Most techniques assume that drift is a function of retention time alone and thus generate a global correction curve $f(rt_A) = rt_B$ (Podwojski *et al.*, 2009). This critical assumption is overly simplistic. In practice, retention time drift is compound dependent (Supplementary Figs S2 and S3) (Smith *et al.*, 2015). Additionally, residual drift

becomes greater when using more vagarious separation strategies such as HILIC, as larger groups of samples are aligned, and as research studies begin to incorporate inter-laboratory comparisons (Abate-Pella *et al.*, 2015).

Given the global correction assumption, most alignment techniques minimize only the average drift between samples considering all analytes equally (Prince and Marcotte, 2006) (Supplementary Figure S2 post correction is an optimistic example of retention drift for technical replicates run over the course of nine hours). As such, the inherent compound-specific drift results in many peaks remaining unaligned after correction—moreover many compounds move even further out of alignment upon global correction (Supplementary Fig. S3). This poor feature alignment causes major challenges for current peak grouping algorithms. The density method employed by XCMS, e.g. can only group peaks if their maximum residual drift is less than the distance to the nearest group (Fig. 1C is an example of this failure) (Smith *et al.*, 2006). Further complexity is added by samples in which a feature is undetected or when spurious noise is detected as a feature, both common occurrences in current workflows.

These failings of the current informatic workflow motivated our development of the Warpgroup algorithm. Warpgroup is an algorithm which utilizes dynamic time warping (DTW) and network graph decomposition. Herein we achieve five goals: (i) accurate grouping of features between samples even in the case of deviation from the global retention time drift, (ii) splitting of peak subregions into distinct groups, (iii) determination of consensus integration bounds within each group such that each group represents a similar chromatographic region, (iv) detection of the appropriate integration region in samples where no peak was detected, and (v) reporting of several parameters which allow filtering of noise groups.

The Warpgroup algorithm establishes a correspondence between the time domains of each feature's extracted ion chromatogram (EIC) trace, utilizing DTW by default (Giorgino; Rabiner, 1978). Based on this correspondence, Warpgroup evaluates whether all supplied peak bounds represent a similar chromatographic region using graph community detection (Csardi and Nepusz, 2006). Subsequently, it determines 'consensus integration regions' for each sample and selects the appropriate integration region for samples with no detected peak. During the time warping and graph analysis, several descriptors of each group are generated and reported for use in filtering unreliable and noise-containing groups.

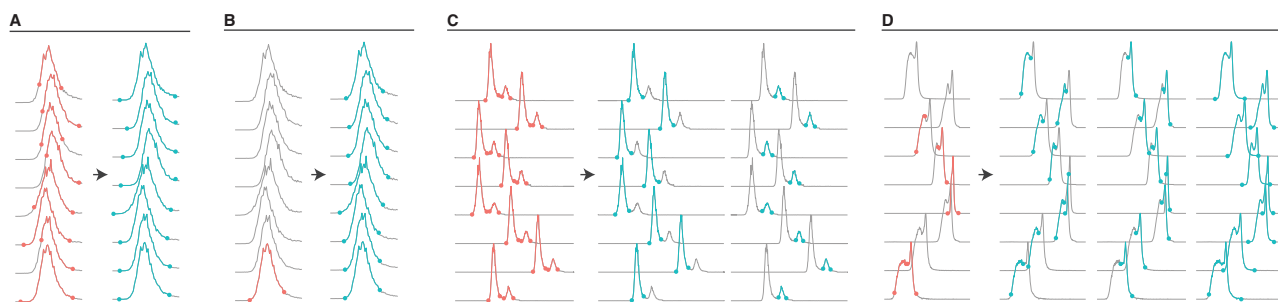


Fig. 1. Examples of conflicts from the traditional workflow which are resolved by Warpgroup. **(A)** Determination of integration bounds is a challenging computational problem. Independent peak detection introduces sample-to-sample variance in the integration regions (left). Peak bounds after Warpgroup (right). **(B)** Peak detection often misses peaks in some samples (left). Warpgroup detects the appropriate regions in each sample to integrate (right). **(C)** Conventional methods are unable to accurately group peaks when retention time varies more than the separation between peaks (left). Warpgroup successfully groups challenging peaks (right). **(D)** An extreme example in which two peaks have merged to varying degrees and peak detection identified different portions of the peak in different samples (left). Warpgroup correctly identifies the three corresponding regions in each sample (right). All examples are included in the Warpgroup R package for demonstration

2 Methods

2.1 Overview of the Warpgroup algorithm

The Warpgroup algorithm is applied after feature detection has been performed. It augments the conventional retention time correction and feature grouping steps with the addition of group splitting and consensus bound determination. The benefits of Warpgroup are derived from the combination of several peak finding rounds by using the independently determined alignment between chromatograms.

The Warpgroup algorithm utilizes two pieces of information. The first piece of information is one EIC trace per sample which includes all of the masses contributing to the peak group. This trace could contain a single detected peak, or multiple peaks per sample depending on the experimental retention time drift and mass drift. These traces are used to determine the pairwise alignment between each sample's time domain for this putative group of compounds. The second piece of information is a list of peak bounds detected in the EIC traces. These must have been determined previously by a peak detection step for at least one sample. The Warpgroup algorithm will use these bounds and the aligned sample traces to split the detected peak list into groups, each of which represent a distinct chromatographic region.

The key assumption made by the Warpgroup approach is that the sample EIC traces exhibit similar topography. Though not strictly true, this is the common assumption made in current retention time alignment techniques (Smith et al., 2015) and has been shown here to be a robust basis for Warpgroup analysis. Under this assumption, we use established methods to warp (shift, expand, and contract) the time domain of the sample EIC trace such that the difference between two sample traces is minimized. In this way we establish a relationship between the two time domains, equating the scans in one sample to the scans in a second for a specific group of compounds (i.e. $f_{m,n}$ (scan in sample m) = scan in sample n). This warping function is taken as the true correspondence between scans in each sample trace and is used to establish relationships between the detected peaks as well as to determine the proper integration region in samples where a peak was not detected.

To this end, the alignment between each sample scan is used to evaluate whether the supplied peak bounds delineate similar or distinct chromatographic regions of their EIC traces. Peak bounds which describe similar chromatographic regions should overlap upon transformation into a second sample's time domain. We ask, for each peak, if the transformed bounds agree. These yes/no answers are expressed as linkages between detected peaks (nodes) creating a graph structure. This graph is split using the walktrap community detection method (Pons and Latapy, 2005) and the resulting communities are taken as peak groups (i.e. groups of peaks which describe similar chromatographic regions).

For each resulting peak group, the full set of transformed peak bounds is then filtered for outliers that do not describe a chromatographic region similar to that of the majority of detected peaks. The mean of the transformed peak bounds within one standard deviation of the mean is taken, defining the 'group-consensus peak bound' for each sample.

Finally, integration bounds must be determined for samples which have no detected peak remaining in the group. It is common for features to be detected in some but not all samples, especially in cases where compounds are of low abundance. Each group's consensus peak bounds are transformed into the missing sample's time domain and the median of these transformed consensus peak bounds is taken as the integration region for the missing sample.

In this way, Warpgroup assures that each peak group contains a region from every sample, each peak group describes a unique chromatographic region, and all peaks in that group describe similar chromatographic regions (Fig. 1).

2.2 Description of the Warpgroup algorithm

2.2.1 Input

The algorithm takes two pieces of information. A sample \times scan sample-trace-matrix (Fig. 1, traces) and a matrix of peak bounds including the peak start, peak end, and sample index (Fig. 1, dots).

2.2.2 Sample trace pre-processing

Optionally, each sample trace is smoothed, padded with 0's equal to 10% of the length of the trace, and normalized to a maximum intensity of 1.0.

2.2.3 Pairwise sample warping matrix generation

Each pair of sample traces is used to generate a sample \times sample warping-matrix (W). Each matrix entry is a step function $W_{m,n} = f_{m,n}(x)$ such that $f_{m,n}$ (scan in sample m) = scan in sample n (Supplementary Fig. S4) The notation $W_{i,j}$ represents the step function converting scans from the sample in which peak j was detected into the sample in which peak i was detected.

The warp matrices for this work are generated using DTW to determine the optimal warp path. Other techniques such as parametric time warping (PTW) have recently been applied to the correction of retention time drift (Wehrens et al., 2015) and in general any technique which establishes alignment between the scans in each sample can be used.

2.2.4 Establishing relationships between the supplied peaks

The supplied peak bounds are transformed from the originating sample's elution space into each of the other sample's elution space via the previously determined warping-matrix. Peaks which delineate the same chromatographic regions will share bounds when transformed from their time domain into the other sample's time domain.

Each pair of peak bounds is compared to populate a peak \times peak match-matrix (P). Pairs which differ by less than the settable cut-off *sc.aligned.lim* are filled as true.

$$P_{i,j} = |\text{bounds}_{\text{peak } i} - W_{i,j}(\text{bounds}_{\text{peak } j})| < \text{sc.aligned.lim}$$

2.2.5 Splitting the supplied peaks into groups which describe distinct chromatographic regions

Matrix P is represented as a graph structure where matrix indices are the nodes and matrix elements containing a true value are the edges (Supplementary Fig. S5). The nodes of this graph are split into communities using the walktrap community detection method (Pons and Latapy, 2005).

Each of the resulting communities contains one or more detected features. Within each community (i.e. group), all detected features are taken to represent the same analyte.

2.2.6 Determination of consensus peak bounds for each group

All detected peaks within each group contribute to the consensus peak bounds such that each peak represents the same chromatographic region. Grouped peak pairs are transformed into each sample's time domain to create a peak \times peak transformed matrix (C).

$$C_{i,j} = W_{i,j}(\text{bounds}_{\text{peak } i})$$

The mean of the bounds within one standard deviation of the mean is taken along each column, defining the consensus peak bounds for the j th peak.

2.2.7 Determination of integration region for samples without a detected peak

For samples in which there was no detected peak remaining in a group, the consensus peak bounds are projected into that sample's time domain.

$$C_{ij} = W_{ij}(\text{consensus bounds}_{\text{peak } i})$$

The median of these transformed bounds are taken as the missing sample's peak bounds.

2.2.8 Output

The output of the algorithm is a list, each entry representing one peak group. A group entry is a matrix with a set of consensus peak bounds for each sample as well as descriptors of the alignment and grouping process. This output can be used for peak integration, filtering, and statistics.

2.3 XCMS implementation

Warpgroup was developed as a standalone algorithm and as such it can be applied to any suitable chromatographic data. For convenience, the Warpgroup package includes integration with XCMS type objects. These functions allow application of the Warpgroup algorithm in the conventional XCMS manner by calling `group.warpgroup()`. The returned object is an `xcmsSet` object with peak bounds and groups generated by the Warpgroup algorithm. This resulting `xcmsSet` does not need any further `fillPeaks()` and is ready for statistical analysis, either manually or with XCMS's `diffreport()` function. Further information can be found in [Supplementary Information page S6](#).

2.4 Datasets

2.4.1 Raw data

We experimentally generated two datasets on which to benchmark Warpgrouping. To evaluate performance under a relevant set of conditions, we chose to generate one dataset with reversed-phase C18 chromatography and the second with amino propyl HILIC (Ivanisevic *et al.*, 2013). Each dataset contained 11 LC/MS runs of *Escherichia coli* (*E. coli*) strain K12, MG1655 metabolic extract. This design allowed us to inspect the standard error of quantitation on both ideal (C18) and non-ideal (HILIC) datasets while also observing the algorithm's performance as dataset quality degrades at longer times.

Metabolic extract was generated as described previously (Mahieu *et al.*, 2014). Briefly, two cultures of *E. coli* were grown, one on natural-abundance glucose and a second on uniformly labeled ^{13}C -glucose as the sole carbon source. *E. coli* was harvested by pelleting 10 ml of culture at $\text{OD}_{600} = 1.0$. Pellets were extracted using 1 ml of 2:2:1 methanol:acetonitrile:water, and reconstituted in 100 μl of 1:1 acetonitrile:water.

Datasets were generated on the Thermo Q-Exactive Plus mass spectrometer interfaced with an Agilent 1260 capillary liquid chromatography system. Spectra were collected with the following HESI II source settings: aux. gas, 15; sheath gas, 30; counter gas, 0; capillary temperature, 310 $^{\circ}\text{C}$; sheath gas temperature, 200 $^{\circ}\text{C}$; spray voltage, 3.2 kV; needle diameter, 34 ga; s-lens, 65 V; mass range, 85–1165 Da; resolution 140 000; microscans, 1; max injection time; 200 ms; automatic gain control target: 3×10^6 .

HILIC was performed as described previously (Ivanisevic *et al.*, 2013) by using the Phenomenex Luna NH2 (1.0 mm \times 150 mm \times 3 μm) column and a flow rate of 50 $\mu\text{l}/\text{min}$. Spectra were collected in negative

ion mode. Solvents were: A, 95% water + 20 mM ammonium hydroxide + 20 mM ammonium acetate; B, 100% acetonitrile. An injection volume of 1 μl was used with a linear gradient of (minutes, %A): 0, 5; 40, 100; 50, 100; 50.5, 40; 54.5, 15; 55, 5; 65, 5.

Reversed-phase chromatography was performed as described previously (Ivanisevic *et al.*, 2013) by using the Agilent Zorbax C18 (0.5 mm \times 150 mm \times 3 μm) column and a flow rate of 30 $\mu\text{l}/\text{min}$. Spectra were collected in positive ion mode. Solvents were: A, water + 0.1% (v/v) formic acid; B, acetonitrile + 0.1% (v/v) formic acid. An injection volume of 1 μl was used with a linear gradient of (minutes, %A): 0, 95; 45, 0; 55, 0; 56, 95; 65, 95.

2.4.2 Pre-processing

The Warpgroup algorithm implements peak subregion detection, consensus/missing peak integration bound determination, and group filtering. These steps come after peak detection has been performed and putative correspondence has been determined. To generate data for comparisons, peak detection for each of the C18 and HILIC datasets was performed by the `centWave` algorithm as implemented in the XCMS R package (R Core Team, 2014; Smith *et al.*, 2006; Tautenhahn *et al.*, 2008). Parameters were: C18, `ppm = 2.5`, `peakwidth = c(8 120)`, HILIC: `ppm = 2.5`, `peakwidth = c(8 120)`. This set of detected peaks was used as the basis for both the conventional and Warpgroup workflows as described below.

2.4.3 Conventional workflow

The conventional workflow as referred to here consists of the following listed analysis steps and parameters taken from XCMS Online recommendations for the Q-Exactive Plus (Tautenhahn *et al.*, 2012). Global retention time correction is performed with the OBI-warp algorithm (`profStep = 1`, `center = 1`). Features are then grouped between samples with the density method (`mzwid = 0.015`). Finally, missing peaks are filled by integrating the range of m/z and retention times in the group using `fillPeaks()`. The resulting filled peak groups contain at minimum one intensity value per sample, but in many instances include multiple intensity values per sample. When performing statistics, the `groupval()` function applies a filter to select a value which will represent each sample. By default, this naively selects the peak which is closest to the median retention time of the group. All calculations are based on this `groupval()` output to make results representative of `diffreport()` output as used in the conventional workflow and by XCMS Online.

2.4.4 Warpgroup workflow

The Warpgroup workflow consists of the following steps. Global retention time correction is performed with the OBI-warp algorithm (`profStep = 1`, `center = 1`). A rough grouping of features is established by grouping all features within 3 ppm and 25 scans. In our data sets this rough grouping ensured that all peaks which could possibly be the same analyte across samples remained in the same group—this also caused some groups to contain multiple peaks. Here, these rough groups were refined with the Warpgroup algorithm by a call to `group.warpgroup` (`rt.max.drift = 20`, `ppm.max.drift = 3`, `rt.aligned.lim = 7`). The resulting dataset contained one peak per sample in every group, all of which described the same region. This output `xcmsSet` was used for all further statistics and assessment of the Warpgroup algorithm.

2.4.5 Selecting peak groups for comparison

The Warpgroup analysis assumes each detected peak represents a legitimate peak region. Upon Warpgroup analysis of these regions, a

single group in the conventional workflow often results in multiple Warpgroup refined groups (which we refer to as warpgroups) (*vide* Table 2). To make a fair comparison between quantitation of the conventional and Warpgroup methods, a selection of groups had to be made. ‘Shared’ groups include any group which contains six or more of the same features across the two workflows. It is worth noting that the ‘shared’ group subset masks the benefits that Warpgroup provides in low abundance peak detection and peak subregion detection.

3 Performance evaluation

A major goal of the Warpgroup algorithm is the reduction of variance introduced by data processing. We identified several points in the conventional bioinformatic workflow for processing untargeted metabolomic data where small errors were being introduced. These small errors were compounded in each downstream step, resulting in a significant decrease in dataset quality. We evaluated the impact of Warpgrouping on the two primary errors we noticed: integration bound selection and peak grouping.

3.1 Peak quantitation

Peak quantitation performance of each workflow was assessed by comparing the coefficient of variation (CV) across 11 replicate injections. This metric provides an assessment of the entire workflow, from peak detection to grouping. Warpgroup often divides conventional groups into multiple sub groups, and thus there is not a one-to-one correspondence between warpgroups and conventional groups. To assess similar peak groups in both methods, only groups sharing more than 6 of the 11 centWave detected peaks were included in the CV analysis. This ensures a one-to-one correspondence between groups from both methods but obscures the benefit of any additional, true groups recovered by Warpgroup.

3.2 Grouping quality

The quality of peak grouping was evaluated for both workflows by manually annotating the resulting groups. Automated rating of group quality is complex and remains beyond the ability of current techniques. To generate a metric for the quality of resulting groups we examined 500 groups generated by each workflow, scoring them for uniformity of included peaks. The scoring system employed was: 4, identical integration regions for every peak in the group; 3, some minor variation in the integration regions; 2, major variation in the integration regions; 1, multiple distinct peaks included in the group; 0, a noise group with no discernable correct integration. Scores were summarized for comparison of the conventional and Warpgroup workflows.

Further, the number of additional, distinct chromatographic regions the Warpgroup algorithm detected was quantified. We manually annotated each Warpgroup as redundant, noise, or unique. Groups that shared >75% of their major chromatographic region, or differed in only the tails of the peak were annotated as redundant. In some cases, a peak was split into subregions but also reported in its entirety (Fig. 1D). We considered this desired behavior and annotated all three as unique groups.

4 Results

4.1 Standard error of replicate injections

Peak picking in the conventional workflow is performed on each sample independently, causing the integration region for each peak to vary slightly from sample to sample. By considering the peak

bounds detected in each sample together, we ensure that a similar integration region is selected for each peak. In addition, the Warpgroup approach reduces errors in grouping which can contribute to inaccurate quantitation and statistics. Analysis of 11 replicate injections with two chromatographies demonstrated the improvement in data processing quality using the Warpgroup method (Fig. 2). The mean CV was halved (a decrease of 13% in the HILIC case and 17% in the C18 case Table 1). Pairwise comparison of each group before and after Warpgroup revealed that, in most but not all cases, the CV decreased with the application of Warpgroup (Supplementary Fig. S7)

4.2 Quality of resulting groups

Grouping in the conventional workflow is based solely on the assumption that common peaks will cluster in retention time. As seen in Supplementary Figures S2 and S3, this assumption is not strictly true and in many cases (such as shown in Fig. 1C) groups will include two or more distinct peaks due to residual drift. We sought to evaluate the quality of peak groups returned by the Warpgroup algorithm as compared with the conventional workflow by rating groups on a scale of 0–4. As seen in Figure 3, the Warpgroup algorithm results in a striking increase in peak group quality.

Notably, ratings of 1 correspond to groups which contained multiple, distinct peaks. The correction of these cases represents an increase in dataset coverage, as the additional groups ‘rescued’ by Warpgroup represent newly quantified unique signals (Fig. 3).

Given that Warpgroup splits peak groups into distinct regions, there is a tradeoff between the number of truly unique chromatographic regions and the number of redundant chromatographic regions that are represented. This tradeoff is controlled by the variable *sc.aligned.lim*, the only user-settable parameter in our algorithm. This parameter specifies the similarity two sets of peak bounds must have to be called the same peak. A smaller *sc.aligned.lim* results in more sensitive peak subregion detection but more orphan peaks and

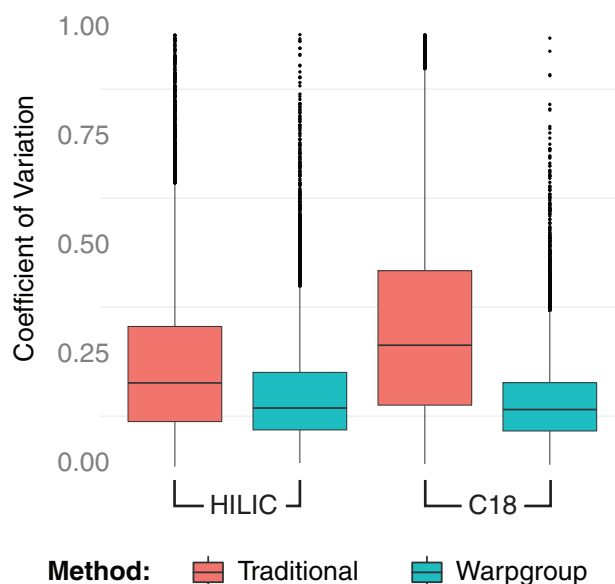


Fig. 2. Standard error of peak quantitation comparison. The CV for all peak groups which shared more than 6 centWave peaks from 11 replicate injections was monitored before (pink) and after (blue) warpgroup. The conventional workflow generates a large number of inconsistent peak groups for various reasons; upon warpgrouping these are corrected, resulting in a much lower CV for the replicates

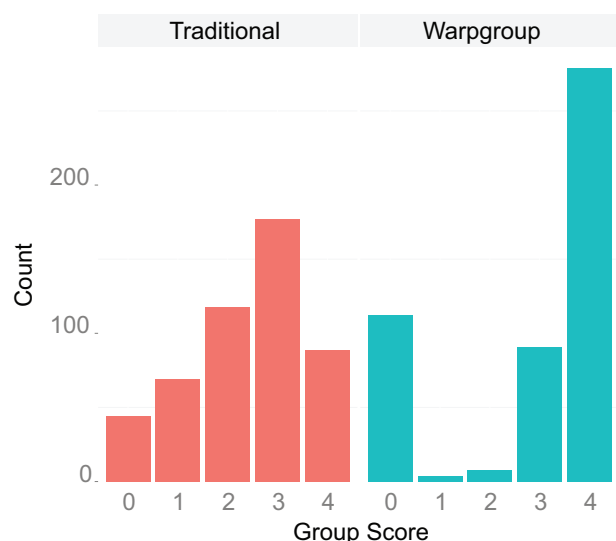


Fig. 3. Group quality and consistency comparison. The conventional XCMS approach without Warpgroup was compared to the XCMS approach with Warpgroup. Quality of generated groups was assessed. Groups were manually inspected and rated on a scale of 0–4. Zero scores corresponded to noise groups with no discernable correct integration. The remaining scores ranged from 1 (integration regions incorporating different peaks across samples) to 4 (identical integration regions across all samples). Warpgroup (right) showed a major improvement in group quality as compared to the conventional workflow (left). Warpgroup also showed an expected increase in noise groups

a larger number of redundant groups. In practice, orphaned and redundant peaks are easily filtered by removing all peak groups generated by one or a few of the originally detected peak bounds.

We evaluated the redundancy of the warpgroups and the increase in unique signals detected by manually annotating redundant and noise-only warpgroups in the HILIC dataset (Table 2). The conventional workflow's 18 341 peak groups resulted in 40 719 peak groups after Warpgrouping. Of these warpgroups, we manually annotated 33% as redundant and 10% as noise. Considering these redundancies and noise, the Warpgroup approach resulted in 23 209 unique signals. The increase in peak groups by 23% represents distinct chromatographic regions that were added to the dataset and otherwise would have been lost or poorly quantified.

A final, more restrictive search for rescued groups was performed. Cases such as that shown in Figure 1C were identified by searching for conventional groups in which two distinct peaks were incorrectly combined by conventional grouping due to residual drift. In these cases, peak finding detected both peaks in most samples but the conventional grouping was unable to properly separate them. This search yielded 611 peak groups in the HILIC dataset and 1246 peak groups in the C18 dataset that were successfully rescued by Warpgroup.

5 Discussion

The exact use cases of Warpgroup are dependent on the data and the problem at hand. We imagine four distinct goals in the next section and summarize appropriate inputs and expected outputs. Warpgroup operates optimally after inter-sample peak correspondence has been established. Though it is possible to supply ungrouped data to the Warpgroup algorithm, there are several drawbacks to this approach. First, processing time for the DTW algorithm scales

Table 1. Coefficient of variation comparisons

Method	Mean CV	90th Percentile CV	Chromatography	Shared peaks
Conventional	33%	57%	C18	15560
Warpgroup	14%	24%		
Conventional	31%	63%	HILIC	7846
Warpgroup	18%	33%		

Table 2. Group quality comparisons

Method	Groups	Percent redundant	Percent noise	Total unique signals
Conventional	18 341	0%	0%	18 341
Warpgroup	40 719	33%	10%	23 209

with the square of the input length. Second, if a feature is present in one sample but missing from the others, this dissimilar topography can result in incorrect alignments. Finally, establishing correspondence is a complex challenge for which many more sophisticated solutions have been suggested (Smith *et al.*, 2015). These should be used in conjunction with the Warpgroup refinements.

Although Warpgroup is not intended to determine peak correspondence, it does make a less restrictive assumption for peak alignment. Current algorithms assume that peak elution order remains monotonic across all masses. Warpgroup assumes only that peaks of indistinguishable mass retain their elution order. This more relaxed assumption allows for rudimentary correspondence to be established in more complex cases such as that shown in Figure 1C and is a major improvement to the XCMS-based workflow (Aberg *et al.*, 2009).

5.1 Use cases

There are four modes we consider for the application of Warpgroup.

5.1.1 Consensus bound determination

In cases where the peak was detected in all samples, only the peak integration region with a small number of surrounding scans need be included in the EIC matrix. A large value for *sc.aligned.lim* can be supplied to avoid splitting of the group into subregions. When operated in this manner, Warpgroup is relatively fast. The returned bounds are the consensus integration bounds for that group (Fig. 1A)

5.1.2 Peak subregion detection

This use case is identical to case one except an appropriately small value for *sc.aligned.lim* is selected, allowing for subregion splitting. This use case is also relatively fast. The returned bounds will be a list of distinct chromatographic regions (Fig. 1D)

5.1.3 Imputation of integration bounds for samples in which no peak was detected

In this case, both the undetected and detected peak traces must be included in the EIC matrix. Because the feature was not detected in at least one sample, the necessary scan range will be dependent on the expected range in which the undetected peak could fall (i.e. the observed drift). A large value for the argument *sc.aligned.lim* should be supplied to avoid splitting the detected peaks into sub groups.

The peak bounds for each missing sample are then returned (Fig. 1B).

5.1.4 Grouping of peaks which deviate from the global retention time drift

Warpgroup's 'grouping' of peaks is a result of the subregion detection and splitting. In this mode, the EIC region supplied to Warpgroup will envelop multiple peak groups and thus take longer than the above modes (Fig. 1C). A small value for *sc.aligned.lim* should be supplied if peak subregion detection is desired, or a larger value if the goal is to distinguish two well-separated peaks. The result will contain a group for each detected peak group.

5.2 Output considerations

One major advantage of the Warpgroup workflow is the ability to detect noise groups. The Warpgroup algorithm includes several peak descriptors for each group after analysis. It is important to note that Warpgroup output is dependent on the input and contains all resulting groups, including noise. Due to the splitting approach, this can result in a large increase in group number—many of which may be redundant. Further, as noise regions have an under-determined warp path, these regions are often split into distinct regions.

Two descriptors generated by the algorithm can be used to detect and filter these cases. These descriptors provide a type of quality measurement of the group. The first descriptor is '*n*'—the number of peaks originally detected which contribute to this group. This parameter is featured in the conventional workflow, but Warpgroup provides a more refined metric. Rather than *n* representing all features eluting near each other, here *n* represents the detected features which describe similar subregions of the chromatogram; thus, the metric is much more reliable. In cases of high *n*, feature detection agreed upon the region of the chromatogram to call a peak. In cases of low *n*, the peak detection did not robustly detect the region and it is likely noise.

The second descriptor is 'warp.consistency'. This metric measures how much the bounds shift when projected into each time-domain and back. Chromatograms with a well-defined and conserved topography will generate highly reproducible warp paths. When bounds are projected through these warp paths, any introduced shift will be small and this metric will be low. When bounds are shifted through a poorly defined region, shifts will be greater and this metric will be high. It is recommended to monitor and filter peak groups based on these parameters prior to further analysis.

5.3 Challenges

A drawback of the Warpgroup approach is speed. As described in Prince *et al.*, 'warping function... [scaling]... is bounded by computational complexity (the more segmented the warp function the more computation required.)' (Smith *et al.*, 2015). Warpgroup segments every distinct mass trace and, as such, the computational demand is high. The DTW algorithm employed scales with the length of the input as $O(n^2)$. Thus, as correspondence confidence decreases, the length of the EIC supplied to Warpgroup increases and processing time lengthens rapidly. Conversely, in cases where correspondence confidence is high or the goal is simply consensus peak bound and subregion detection for well-grouped peaks, the algorithm remains very fast. Accordingly, the incorporation of mass and retention time drift correction as well as the establishment of correspondence prior to Warpgroup is recommended.

Prince *et al.* raise several limitations of current correspondence methods (Smith *et al.*, 2015). Though not intended as a correspondence algorithm, Warpgroup does address some of the challenges

these methods face. Most importantly, Warpgroup makes more realistic assumptions about the component-specific drift expected in these datasets. Further, as a single reference sample is not used for alignment, Warpgroup remains symmetric and robust. The algorithm is easily implemented in most workflows as it relies on only one required user-settable parameter.

5.4 Future directions

Improving the scaling with sample number is an important goal. Although the current implementation is sufficient for many published metabolomic studies, the analysis of larger datasets remains a priority. Computation can be minimized by several strategies. For many peak groups, refinement with Warpgroup will be unnecessary, making minor or no modification to the predetermined group. In these cases, Warpgroup can be omitted for all but the most complex groups. The major computational step is the establishment of a warp path between each sample. To reduce computation, the DTW algorithm can be replaced with faster warping algorithms such as PTW if the data allow. Finally, this implementation calculates the full sample \times sample warping matrix. However, implementation of a sparse matrix approach could be explored.

Although Warpgroup was presented here in the context of LC/MS data, the input and output of the algorithm are of a general form (multiple time series and regions within those time series.) As such, the method is generalizable and can find consensus regions within any time-series data. An example of Warpgrouping on echocardiogram data (Goldberger *et al.*, 2000; Penzel *et al.*, 2000) can be found in Supplementary Figure S9.

The Warpgroup algorithm as presented addresses several major drawbacks of the current informatic workflow. Still, current processing techniques leave significant room for improvement. The development of more effective correspondence algorithms is a critical step for the advancement of the field (Smith *et al.*, 2015). Additionally, we see promise in leveraging the information embedded in the component-specific drift observed in these datasets. For example, the drift data may be used to cluster ions into composite spectra and to inform further identification.

6 Conclusion

In summary, we have found Warpgroup to be an important refinement step for current integration and correspondence methods. With Warpgroup refinement in place, data processing results remain robust across a wide range of experimental conditions. Major advantages have been noted in coverage as well as quantitation, especially in low-abundance signals. Further, Warpgroup output includes additional descriptors which can be used to filter noise and unreliable groups from the final datasets. Overall, we expect the addition of a Warpgrouping step to the informatic workflow to improve the quality and reliability of untargeted metabolomic analyses.

Supporting data

The LC/MS datasets used in benchmarking of the Warpgroup algorithm can be found on our laboratory website at <http://pattilab.wustl.edu/software/warpgroup/>.

Acknowledgements

We would like to thank Dr. Robert Pless for his valuable guidance in helping develop this work.

Funding

G.J.P. received financial support for this work from the National Institutes of Health Grants R01 ES022181 and L30 AG0 038036, as well as the Alfred P. Sloan Foundation, the Camille and Henry Dreyfus Foundation, and the Pew Scholars Program in the Biomedical Sciences.

Conflict of Interest: none declared.

References

- Abate-Pella, D. *et al.* (2015) Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods. *J. Chromatogr. A*, **1412**, 43–51.
- Aberg, K.M. *et al.* (2009) The correspondence problem for metabonomics datasets. *Anal. Bioanal. Chem.*, **394**, 151–162.
- Buszewski, B. and Noga, S. (2012) Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique. *Anal. Bioanal. Chem.*, **402**, 231–247.
- Cappadona, S. *et al.* (2012) Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino Acids*, **43**, 1087–1108.
- Crutchfield, C.A. *et al.* (2010) Mass spectrometry-based metabolomics of yeast. *Methods Enzymol.*, **470**, 393–426.
- Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal Complex Sy*, 1695.
- Fuhrer, T. and Zamboni, N. (2015) High-throughput discovery metabolomics. *Curr. Opin. Biotechnol.*, **31**, 73–78.
- Giorgino, T. (2009) Computing and visualizing dynamic time warping alignments in R: the dtw package. *J. Stat. Softw.*, **31**, 1–24.
- Goldberger, A.L. *et al.* (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, **101**, e215–e220.
- Ivanisevic, J. *et al.* (2013) Toward 'omic scale metabolite profiling: a dual separation–mass spectrometry approach for coverage of lipid and central carbon metabolism. *Anal. Chem.*, **85**, 6876–6884.
- Käll, L. and Vitek, O. (2011) Computational mass spectrometry-based proteomics. *PLoS Comput. Biol.*, **7**, e1002277.
- Kele, M. and Guiochon, G. (2000) Repeatability and reproducibility of retention data and band profiles on reversed-phase liquid chromatography columns. *J. Chromatogr. A*, **869**, 181–209.
- Mahieu, N.G. *et al.* (2014) Credentialing features: a platform to benchmark and optimize untargeted metabolomic methods. *Anal. Chem.*, **86**, 9583–9589.
- Nikolskiy, I. *et al.* (2013) An untargeted metabolomic workflow to improve structural characterization of metabolites. *Anal. Chem.*, **85**, 7713–7719.
- Patti, G.J. *et al.* (2012) Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.*, **13**, 263–269.
- Penzel, T. *et al.* (2000) The apnea-ECG database. In: *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*. IEEE, pp. 255–258.
- Podwojski, K. *et al.* (2009) Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics*, **25**, 758–764.
- Pons, P. and Latapy, M. (2005) Computing communities in large networks using random walks (long version). *Lect. Notes Comput. Sci.*, **3733**, 284–293.
- Prince, J.T. and Marcotte, E.M. (2006) Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.*, **78**, 6140–6152.
- Quarry, M.A. *et al.* (1984) Measurement and use of retention data from high-performance gradient elution. *J. Chromatogr. A*, **285**, 19–51.
- R Core Team (2014) R: A Language and Environment for Statistical Computing.
- Rabiner, L.R. (1978) Considerations in dynamic time warping algorithms for discrete word recognition. *J. Acoust. Soc. Am.*, **63**, S79.
- Smith, C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.
- Smith, R. *et al.* (2015) LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Brief. Bioinform.*, **16**, 104–117.
- Tautenhahn, R. *et al.* (2008) Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, **9**, 504.
- Tautenhahn, R. *et al.* (2012) XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.*, **84**, 5035–5039.
- Vandenbogaert, M. *et al.* (2008) Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics*, **8**, 650–672.
- Wehrens, R. *et al.* (2015) Fast parametric time warping of peak lists. *Bioinformatics*, **31**, 3063–3065.