BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 19 2013, pages 2469–2476
doi:10.1093/bioinformatics/btt435

Bioimage informatics

Advance Access publication July 30, 2013

Graph-based peak alignment algorithms for multiple liquid chromatography-mass spectrometry datasets

Jijie Wang[1] and Henry Lam[1,2,*]

[1]Division of Biomedical Engineering and [2]Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Associate Editor: Jonathan Wren

ABSTRACT

Liquid chromatography coupled to mass spectrometry (LC-MS) is the dominant technological platform for proteomics. An LC-MS analysis of a complex biological sample can be visualized as a 'map' of which the positional coordinates are the mass-to-charge ratio (m/z) and chromatographic retention time (RT) of the chemical species profiled. Label-free quantitative proteomics requires the alignment and comparison of multiple LC-MS maps to ascertain the reproducibility of experiments or reveal proteome changes under different conditions. The main challenge in this task lies in correcting inevitable RT shifts. Similar, but not identical, LC instruments and settings can cause peptides to elute at very different times and sometimes in a different order, violating the assumptions of many state-of-the-art alignment tools. To meet this challenge, we developed LWBMatch, a new algorithm based on weighted bipartite matching. Unlike existing tools, which search for accurate warping functions to correct RT shifts, we directly seek a peak-to-peak mapping by maximizing a global similarity function between two LC-MS maps. For alignment tasks with large RT shifts (>500 s), an approximate warping function is determined by locally weighted scatterplot smoothing of potential matched features, detected using a novel voting scheme based on co-elution. For validation, we defined the ground truth for alignment success based on tandem mass spectrometry identifications from sequence searching. We showed that our method outperforms several existing tools in terms of precision and recall, and is capable of aligning maps from different instruments and settings.

Availability: Available at https://sourceforge.net/projects/rt-alignment/.

Contact: kehlam@ust.hk

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

Received on April 16, 2013; revised on July 22, 2013; accepted on July 24, 2013

1 INTRODUCTION

Liquid chromatography-mass spectrometry (LC-MS) is the dominant technological platform in proteomics, capable of profiling thousands of proteins in the biological sample in a high-throughput and sensitive manner. In the most popular workflow of shotgun proteomics, proteins in a complex mixture are first digested into small peptides, which are then separated by high-performance liquid chromatography, and eluted through an electrospray ion source into the mass spectrometer. In the mass spectrometer, the peptide ions are separated by their mass-to-charge ratios (m/z) and converted to an electrical signal, the intensity of which is a measure of abundance of the peptide ion. Taken together, LC-MS data are three-dimensional, with each peptide ion represented by three coordinates: its chromatographic retention time (RT), its m/z and its signal intensity. In some instruments, an additional step of tandem mass spectrometry (MS/MS or MS2) can be taken to fragment each peptide, yielding fragment ion (MS2) spectra from which the peptide can be identified by computational methods. An LC-MS 'map' can therefore be viewed as a comprehensive and quantitative profile of the proteome being analyzed. Many proteomics experiments rely on comparisons of LC-MS maps of different samples (Bisle *et al.*, 2006; Böttcher *et al.*, 2008; Catchpole *et al.*, 2005; Vissers *et al.*, 2007). For these experiments, it is necessary to match the peak originating from the same peptide in different LC-MS maps. This task can be easily accomplished if the m/z and the RT coordinates of the same peptide stay constant in many LC-MS runs. The m/z values usually vary within a limited and predictable range, which depends on the mass resolution of the mass spectrometer. However, the RT of a specific peptide heavily depends on the instrument conditions and the composition of the mixtures. For many technical reasons, even the same sample analyzed by the same instrument with the same settings can generate different LC-MS maps with shifts and distortions in the RT dimension. This fact makes the association of the same peptides across LC-MS maps difficult because the shift of any two LC-MS maps cannot be known in advance. To find the corresponding peptides in the different LC-MS maps, time alignment must be conducted.

The ultimate goal of time alignment is to match peaks that belong to the same peptides across different LC-MS maps. Based on the type of input data on which the alignment algorithm operates, LC-MS alignment methods can be categorized into 'profile-based' methods and 'feature-based' methods. Profile-based methods directly process the raw data obtained in the LC-MS experiment, aiming to avoid errors introduced by a feature-finding algorithm. The early profile-based methods compare only the difference in total ion chromatograms (TIC), which is the sequence of the sum of all signals at each time point. Typical methods used in aligning TIC include dynamic time warping (DTW) (Sakoe and Chiba, 1978), correction optimized warping (COW) (Nielsen *et al.*, 2002), parametric time warping (PTW) (Eilers, 2004) and continuous hidden Markov models (HMM) (Listgarten *et al.*, 2005). TIC-based methods make little use of

*To whom correspondence should be addressed.

the whole LC-MS map, disregarding most of the useful information at the m/z dimension that can help alignment. Therefore, recent approaches tried to use more information available in the LC-MS maps by applying these time warping methods to many isolated ion chromatograms at different m/z values (Christin *et al*., 2008, 2010; Listgarten *et al*., 2007; Suits *et al*., 2008; Wang *et al*., 2003) or by defining the spectral similarity of corresponding mass spectra as the objective function to be maximized (Finney *et al*., 2008; Prakash *et al*., 2006; Prince and Marcotte, 2006). In general, profile-based methods can produce more accurate alignment results, but they are computationally expensive and may require more resources in terms of computer cycles and memory, especially for multiple alignments (Lange *et al*., 2007, 2008; Vandenbogaert *et al*., 2008). Another limitation of the aforementioned methods is that they only allow shifting, stretching or shrinking of the time axis, and cannot deal with the reversal of elution order. A recent publication suggested that elution order reversals are not rare, especially at fine timescales (Snyder and Dolan, 2006), and when the LC-MS maps are acquired in slightly different settings.

Feature-based approaches operate on 'features' in the LC-MS map detected by a feature-finding algorithm, which tries to distinguish between peptide features and irrelevant noise in the LC-MS experiment. The resulting feature list, which condenses the raw data into (m/z, RT) coordinates of interest, are used in the subsequent alignment. Many feature-based alignment algorithms have been developed, either implemented as stand-alone tools or as a part of software platforms for label-free quantitative proteomics. Examples are MZmine (Katajamaa and Orešič, 2005), MapAligner in OpenMS (Lange *et al*., 2007), SuperHirn (Mueller *et al*., 2007), XCMS (Smith *et al*., 2006), SpecArray (Li *et al*., 2005) and XAlign (Zhang *et al*., 2005). Except MZmine, which simply attempts to find peak matches within a tolerance window, all of these methods attempt to estimate a linear (MapAligner, XAlign) or nonlinear warping function (others) to correct the RT axis. More recently, Zhang (2012) has proposed a divide-and-conquer algorithm that can handle both low-resolution and high-resolution LC-MS maps. Noy *et al*. (2011) propose a novel feature-matching algorithm using a robust nonparametric kernel-type regression model to detect a nonlinear alignment, and a wavelet-based method to incorporate peak shape information for resolving ambiguous matches. In contrast to profile-based methods, feature-based approaches typically rely on high-resolution data for effective feature finding, whereas profile-based methods are generally able to work well on low-resolution LC-MS data. Time alignment methods for LC-MS data are reviewed in Åberg *et al*. (2009), America and Cordewener (2008), Dowsey *et al*. (2010), Lange *et al*. (2008), Katajamaa and Orešič (2007) and Vandenbogaert *et al*. (2008).

Accurate mass and time (AMT) tagging is another related class of LC-MS data processing methods (Jaitly *et al*., 2006; May *et al*., 2007). In AMT methods, the objective is to associate each peptide with the m/z and RT coordinates at which it is expected to be found, such that peaks detected in future LC-MS can be identified to the originating peptides without the help of MS/MS. To correct for potential RT shifts caused by instrument noise or changes in experimental conditions, AMT methods also need to perform time alignment, but in a different manner. Given the peptide identifications, AMT methods typically attempt to associate the raw RT with a predicted hydrophobicity measure of the peptide, and store these in an AMT database, usually specialized to each biological sample. Then, subsequent LC-MS maps from the same sample are analyzed by first converting raw RT into a hydrophobicity measure, usually using information of the elution gradient program, and then searching the coordinates in the AMT database to find the closest matched peptide. AMT methods are therefore theoretically capable of aligning LC-MS maps across vastly different conditions, as time alignment is assisted by the knowledge of the peptide identification and the elution gradient. Potentially large RT shifts caused by differences in elution conditions can be corrected. However, AMT methods depend on the availability of peptide identifications and the accuracy of hydrophobicity predictions, and may not be generally applicable to all types of LC-MS data.

In this article, we propose a feature-based LC-MS alignment method called LWBMatch, based on maximum weighted bipartite matching. LWBMatch is a multiple alignment method that is capable of aligning maps with large systematic and nonlinear RT shifts. Similar to other feature-based methods, an optional preliminary step first extracts likely matching features in the LC-MS maps to detect potential large RT shifts. A warping function is then fitted by locally weighted scatterplot smoothing (LOWESS). However, unlike existing methods that try to determine the warping function accurately on a fine timescale, our warping function needs only to be approximate and mainly serves to bring matching features into proximity. The key innovation of our method is the subsequent step that relies on a global maximization algorithm based on weighted bipartite matching to match features exhaustively, producing an optimal feature-to-feature mapping that maximizes the similarity between two LC-MS maps. As a result, small elution order changes of the peptides and other RT variations caused by instrument noise will have minor effects on the performance. We demonstrate that our method can align LC-MS maps generated from the same conditions even without the warping step, and can also handle maps generated from different instruments and settings after approximate warping. We also show that our method outperforms three existing feature-based alignment methods.

2 MATERIALS AND METHODS

2.1 Evaluation datasets

In our article, we attempted to align LC-MS maps with different level of similarity. We approximately categorized alignment tasks into two groups. 'Homogeneous' alignment is performed on datasets taken on the same LC-MS instrument under identical conditions, within the same batch of experiments. 'Heterogeneous' alignment is performed on datasets from different instruments, under different chromatographic conditions, or separated by long time intervals between acquisitions.

2.1.1 The Standard Protein Mix Database The Standard Protein Mix Database (Klimek *et al*., 2007) is a dataset generated by performing repeated analyses of a standard sample of 18 trypsin-digested proteins using 8 different mass spectrometers. Each analysis consists of 10 consecutively run replicates using the same chromatography column on the same machine. Alignment of these replicates is 'homogeneous' by our definition. Five standard mix preparations (Mix 1, Mix 2, Mix 3, Mix 4 and Mix 7) are analyzed at different times over many years. Alignment

of LC-MS maps from different 'Mix' datasets is 'heterogeneous'. To validate our alignment, the peptide identifications of the MS2 spectra, determined by SEQUEST (Eng *et al.*, 1994), are used as ground truths. Details about sample preparation, mass spectrometry setting and searching parameters can be found in Klimek *et al.* (2007). In our evaluation, we selected one analysis generated from high-resolution instruments per 'Mix' (LTQ-FT for Mix 1, QSTAR for Mix 2, QTOF for Mix 3 and Mix 4, LTQ-Orbitrap for Mix 7) as the test datasets. This is because the feature-finding method we used in OpenMS, like other existing feature-finding methods, only works well on high-resolution datasets.

2.1.2 U2OS Cell Dataset To test our method on a dataset with whole proteome complexity, the U2OS Cell Dataset (Geiger *et al.*, 2012) was used. The U2OS Cell Dataset contains triplicate analysis of a human cell line U2OS, each containing six runs from different peptide fractions. One replicate was acquired on an LTQ-Orbitrap Velos mass spectrometer, while the other two replicates were run in a separate LTQ-Orbitrap Elite mass spectrometer. For the latter, the two replicates were also acquired at different resolution and scan rates. In this dataset, MS/MS spectra were searched by the Andromeda search engine. Further information about sample preparation, protein digestion and fractionation, LC-MS analysis and MS/MS searching can be found in Geiger *et al.* (2012). Alignment of these replicates is treated as 'heterogeneous' because of the difference in instrument or acquisition settings.

2.1.3 Feature detection Feature detection is not the focus of this study. To conduct a fair evaluation, we detected and extracted peptide signals in all the raw data maps using FeatureFinder, which is one of the tools in the OpenMS suite (Kohlbacher *et al.*, 2007), as inputs for all the alignment methods. Therefore, notwithstanding potential misses or errors in the feature-finding step, all alignment methods are compared on a fair ground. A recent publication has compared many different feature detection methods and found OpenMS's FeatureFinder to be one of the best in its class (Hoekman *et al.*, 2012).

2.2 WBMatch and homogeneous alignment

Our homogeneous alignment method, called WBMatch, is inspired by maximum weighted bipartite matching. Unlike existing feature-based alignment methods, it makes no assumption on the warping function of the RT axis. Instead, we transform the homogeneous alignment problem to a weighted matching problem in bipartite graphs and try to work out an optimal matching. In the first step, pairwise weight computations are made for each pair of features (r, s) with $r \in reference$ and $s \in sample$. The weight function w is defined in Section 2.3. In the second step of WBMatch, a well-known weighted bipartite matching algorithm called Kuhn–Munkres algorithm (Kuhn, 1955; Munkres, 1957) will be applied. Because our adjacency matrix is sparse, an improvement using Fibonacci heaps for the shortest path computation is adopted (Fredman and Tarjan, 1987). The complexity of the algorithm is $O(|U|^2 log(|U|))$, where $|U|$ is the number of features to be matched in the larger set. The output of the Kuhn–Munkres algorithm is a mapping that attains the maximum total weight $W = \sum w$, summed over all matched features. Note that the algorithm attempts to find a unique partner in the second map for each feature in the first map. After the maximization, paired features with a weight of 0 in the final matching, which indicate that they are outside of the tolerance window, are considered unmatched. The flowchart and pseudocode of WBMatch are given in Supplementary Figures S1 and S2, respectively. More details about maximum weight bipartite matching can be found in a comprehensive graph theory textbook (West, 2001).

2.3 Determination of the weight function

The weight function is used to measure the similarity between two features. We first define the distance between two features as follows:

$$D(r,s) = \begin{cases} \infty, \text{ if } d_{RT} > \Delta_{RT} \text{ or } d_{MZ} > \Delta_{MZ} \\ \frac{1}{\Delta_{RT}} d_{RT} + \frac{1}{\Delta_{MZ}} d_{MZ}, \text{Others} \end{cases} \quad (1)$$

$$d_{RT} = |RT(r) - RT(s)|, d_{MZ} = |MZ(r) - MZ(s)| \quad (2)$$

where d_{RT} and d_{MZ} refer to the deviations in RT and m/z between two features (r, s) in one pair. Δ_{RT} and Δ_{MZ} are the maximum tolerances for d_{RT} and d_{MZ}, respectively. The distance for any pair with difference in RT or m/z larger than the tolerance will be assigned to a score of ∞. We tested four different transformation functions for converting the distance measure into a similarity measure (Supplementary Fig. S3), and selecting the fourth one that represents the best balance between precision and recall:

$$w(D) = \left\lfloor \frac{1000}{1 + e^{\mu \cdot D - \lambda}} \right\rfloor \quad (3)$$

where we set $\mu = 18$ and $\lambda = 6$. This logistic function-like transformation implements a 'soft' cutoff: as the distance grows, the similarity decreases slowly at first to retain good-quality matching pairs, and then decreases sharply to remove the low-quality matching pairs. A three-dimensional visualization of weight function is shown in Supplementary Figure S4. All parameters are optimized using the Standard Protein Mix Database.

2.4 LWBMatch and heterogeneous alignment

In heterogeneous alignment, LC-MS maps acquired on different instruments and under different settings are aligned. For such maps, even though the time dimension can be shifted and distorted by large amounts, sometimes nonlinearly, the RT deviations for corresponding features across maps are still expected to follow a smooth curve. This is a necessary consequence of the fact that elution orders are approximately conserved. Given this observation, we combined LOWESS and WBMatch to tackle the problem of heterogeneous alignment, which we called LWBMatch. The flowchart and pseudocode are given in Supplementary Figures S1 and S5, respectively. The LOWESS procedure, which will produce an approximate warping function between two maps, consists of two steps. First, we extracted candidate feature pairs from different maps if their m/z difference is within Δ_{MZ}. Using only m/z tolerance will yield many incorrectly matched candidates from which to fit the warping function. To identify the correct feature pairs among the incorrect ones, we applied a voting strategy, shown in Supplementary Figure S6. Based on the simple idea that co-eluting peptides in one map will likely also be co-eluting in the other, for each time point in the reference map, we selected the time point along the RT axis in the sample map with the largest numbers of common features. Second, given these correct feature pairs, the robust locally weighted regression model (Cleveland, 1979) is applied on all the potential RT deviations to yield an affine transformation as the desired warping function (Supplementary Fig. S7A). This warping function is applied to correct the RTs in one map to match those of the other map. After this correction, most of the new RT deviations are within 500 s (Supplementary Fig. S7B), and thus WBMatch can be called on to produce more accurate feature-to-feature mapping.

2.5 Multiple alignment and hierarchical clustering

Multiple alignment, which is the alignment of more than two LC-MS maps, is often necessary in larger studies. Pairwise alignment of all maps in the set is often infeasible, as the computational expense will grow exponentially with the number of maps. Therefore, LWBMatch constructs a hierarchical alignment topology based on the similarity of

LC-MS maps automatically. For every possible pair of maps, the approximate measure of similarity is calculated as the average RT shifts of all potential feature pairs detected before the LOWESS step (defined in Section 2.4). Smaller average shifts indicate that the two maps are more similar, and more similar maps will be aligned earlier in a hierarchical clustering strategy.

3 EVALUATION STUDY

3.1 Ground truth, precision and recall

We established ground truth for the Standard Protein Mix Database and U2OS cell datasets by means of MS/MS information and identification results. Our ground truth consists of features that could be annotated with reliable peptide identifications. PeptideProphet and iProphet (Keller *et al*., 2005; Shteynberg *et al*., 2011) were used to assess the reliability of peptide identifications, and we retained identifications at a false discovery rate of 1%. Then, in each LC-MS map, whenever present, multiple MS/MS spectra that are assigned to the same identification are grouped together and represented by a single feature of coordinates equal to the average of those of the replicate MS/MS spectra. The ground truths from different maps to be aligned are listed in an identification matrix, where each row is a unique identification, and each column contains the features of one LC-MS map. An example of identification matrix is shown in Supplementary Table S1. In each row, every pair of two features is counted as a ground truth. For example, if a given identification can be found in 5 maps, then there are 10 $(= 4+3+2+1)$ feature pairs in the ground truth set that can be potentially aligned.

The purpose of LC-MS alignment is to find groups of corresponding features that originate from same peptides. From this perspective, we formulate the measure of alignment success as that of an information retrieval problem. In information retrieval, 'precision' is the fraction of retrieved instances that are relevant, whereas 'recall' is the fraction of relevant instances that are retrieved. In our case, the relevant instances are feature pairs across different maps with the same peptide identification, whereas the retrieved instances are the feature pairs found by our algorithm. A perfect alignment will have both *precision* and *recall* equal to 1. False positives (erroneously matched features) will lower the alignment precision and false negatives (erroneously unmatched features) will lower the alignment recall. For each dataset, we calculated these performance metrics for our method and three different benchmarks described below, with parameters listed in Supplementary Table S2. For the purpose of calculating recall, we ignored all identifications whose coordinates are not located close to any features in the map. These omissions are likely due to the imperfections in the feature-finding algorithm.

Note that the size and identification accuracy of the ground-truth set is influenced by the choice of the false discovery rate cutoff, which is arbitrary, albeit customary, in shotgun proteomics experiments. Therefore, the precision and recall values are intended to be metrics for the comparison of different methods given identical input. Also note that in our testing, we allow the program to align all maps in the Standard Protein Mix datasets and U2OS datasets as a multiple alignment problem, regardless of whether the alignment is homogenous (within dataset) or heterogeneous (across dataset). To highlight the performance specifically for the more difficult heterogeneous alignment, we only count pairs that are present across at least two different datasets in the calculation of precision and recall.

3.2 Estimating the difficulty of the alignment task

We used two metrics to estimate the difficulty of a given alignment task. The first is the distribution of RT shifts of ground-truth pairs. Larger average RT shifts and larger variance in the RT shifts imply more difficult alignment tasks owing to systematic and random RT variations that likely result from changes in experimental conditions. The second is the frequency of elution order reversals. The more frequently pairs of peptides elute in opposite orders in different maps, the more difficult the alignment task should be. To count the elution order reversals, ground-truth pairs are first sorted by their RTs in the reference map (RT_1), from smallest to largest. If the elution order is conserved, the RT of the corresponding feature in the other map (RT_2) should be monotonically increasing as we go down this list. We count the number of times that the rank of RT_1 increases in the sorting list while the rank of RT_2 decreases; this is the number of elution order reversals between those two maps. To allow for some inevitable noisy variations in locating the apex of the elution peak by the feature-finding algorithm, we round the RT to the nearest 50 s in this exercise, so that these small variations will not be counted as elution order reversals. We report the ratio of the number of elution order reversals to the total number of ground-truth pairs (Fig. 1).

3.3 Benchmarking against existing alignment methods

Three existing feature-based alignment methods were used as benchmarks for our method: the tool MapAligner (Lange *et al*., 2007) in OpenMS (Sturm *et al*., 2008), SuperHirn (Mueller *et al*., 2007) and a modified DTW implemented by us. For all alignment tasks, we applied LWBMatch rather than WBMatch because in the case of homogenous alignment, LWBMatch will degenerate into WBMatch because the LOWESS model will generate a warping function close to the diagonal, such that essentially no warping is applied.

All tools are challenged with a multiple alignment problem comprising all LC-MS maps of the dataset(s), with the features in each map detected by the same FeatureFinder algorithm in OpenMS beforehand. Note that the MapAligner and the modified DTW use the map with the most features as the starting point and align all other maps to this one. Superhirn uses a hierarchical clustering approach similar to ours, albeit with different definitions of the similarity measure between pairs of LC-MS maps. Briefly speaking, the similarity measure in SuperHirn is based on the overlap of features and reproducibility of their intensity values, whereas ours is based on the potential RT shifts.

For a fair comparison, LWBMatch, MapAligner and SuperHirn are tested using the same matching tolerance windows (Supplementary Table S2). Note that all three tools use two sets of tolerances. The pre-warping tolerance specifies how far one can go to find matching 'landmark' feature pairs as input to fitting the warping function. To perform heterogeneous alignment with potentially large and unpredictable RT shifts, only an m/z tolerance based on the mass resolution of the instrument

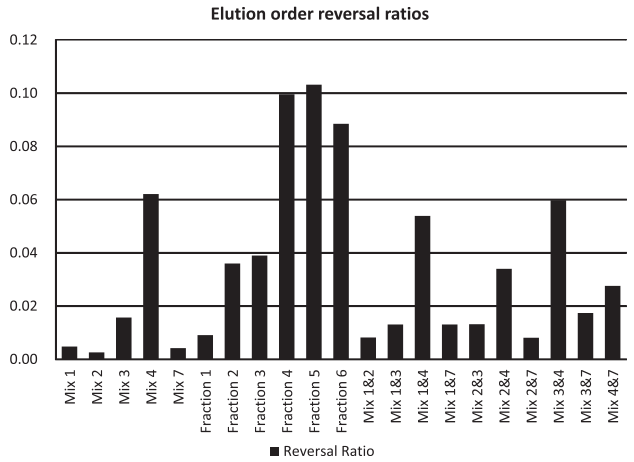

Fig. 1. The elution order reversal ratios of different evaluation datasets

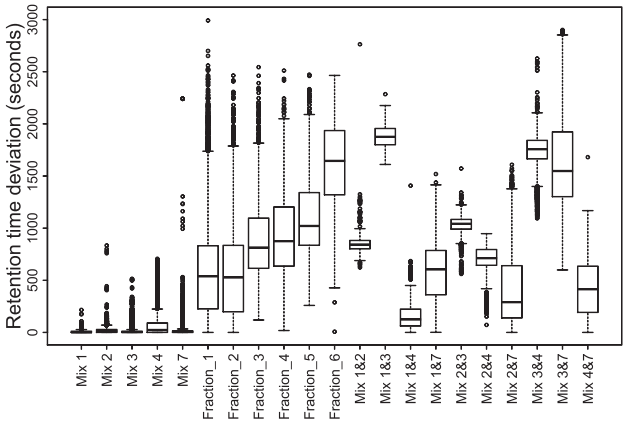

Fig. 2. The boxplot of RT deviations. The bottom and top of the box are the first and third quartiles, and the band inside the box is the second quartile. The ends of the whiskers below and above the box mean the lowest and highest datum within 1.5 Interquartile Range (IQR) of each quartile, respectively. Any data not included between the whiskers are plotted as an outlier with a small circle

is used. No restriction is placed on the RTs of the matching feature pair for all three tools. The post-warping tolerance specifies the maximum deviations within which one can declare two features to be matched, after applying the warping function. For LWBMatch, this is the tolerance used in the WBMatch step, i.e. Δ_{RT} and Δ_{mz}.

DTW is originally a signal-based method, but in this study we adapted it to work on feature maps. Features in the same map are sorted by RT from lowest to highest, resulting in a sequence of feature vectors. Dynamic programming is applied on the two sequences to be aligned to maximize the similarity of the two sequences by allowing shifts. The score of the alignment path P is calculated iteratively in equation 4. A diagonal transition can be performed on Cond. 2, whereas a shrink and a stretch in *reference* relative to *sample* can be conducted on Cond. 1. The action with highest score of corresponding path will be taken. In the end, the highest scoring path $P(n, m)$, where n and m refer to the number of features in *reference* and *sample*, respectively, specifies how features in one map should be matched to those in the other.

$$P(i,j) = max\begin{cases} max\{P(i-1,j), P(i,j-1)\}, \text{Cond. 1} \\ P(i-1,j-1)+1, \text{Cond. 2} \end{cases} \quad (4)$$

$$\begin{cases} \text{Cond. 1} : |\text{MZ}(reference_i) - \text{MZ}(sample_j)| > MZ_t \\ \text{Cond. 2} : |\text{MZ}(reference_i) - \text{MZ}(sample_j)| \leq MZ_t \end{cases}$$

For DTW, there is no separate step of finding landmark feature pairs to fit an approximate warping function. Therefore, only an m/z tolerance is specified (MZ_t), which is set equal to Δ_{mz}. There is no restriction on the RT shift allowed.

The running time performance of LWBMatch, as compared with the benchmarks, is shown in Supplementary Figure S8.

4 RESULTS

4.1 Observed RT shifts and elution order reversal of ground truth pairs

We first estimate the difficulty of the alignment tasks in terms of RT shifts and elution order reversals. To visualize the distribution of RT deviations for these datasets, the absolute values of the RT deviations of all ground-truth pairs are shown in a box-whisker plot (Fig. 2). Most of the RT shifts in the homogeneous datasets are <200 s, except in Mix 4, in which one-fifth of the RT shifts are >200 s. This can be traced to two runs: QT20060925_mix4_23 and QT20060926_mix4_19, in which peptides appear to elute ~400 s later than in the other eight runs. Compared with the RT shifts of homogeneous datasets, the RT shifts in the Standard Protein Mix and U2OS datasets are much larger, which range from 200 s to 2000 s. In addition, the boxes' widths is wider, implying that the RT time shifts also vary more significantly.

The frequency of elution order reversal for all the test datasets is shown in Figure 1. From the figure, the elution order reversal ratios of homogeneous datasets are <2% except Mix 4. This is one important reason why the alignment performance for Mix 4 is not so good compared with that of other datasets, shown in Figure 3A. For the heterogeneous cases, the average of elution order reversal ratios of U2OS datasets and the Standard Protein Mix datasets are 6.3 and 2.4%, respectively, which is larger than that of homogeneous cases.

4.2 Results for homogeneous alignment

To test the performance of multiple homogenous alignment, we selected 10 consecutively run replicates on the same machine from the Standard Protein Mix dataset. This includes the LTQ-FT runs of Mix 1, QSTAR runs of Mix 2, QTOF runs of Mix 3, QTOF runs of Mix 4 and the LTQ-Orbitrap runs of Mix 7. In addition, OR20070924_S_mix7_11 from Mix 7 was removed because this map has a little signal and appears to be a failed LC-MS run.

Figure 3A shows the results for different homogeneous alignment methods. The performance of the LWBMatch is among the best, especially for the challenging dataset Mix 4, which has by far the largest RT shifts and elution order reversal ratios among those tested. Generally, LWBMatch offered similar precision to other methods, but much higher recall values,

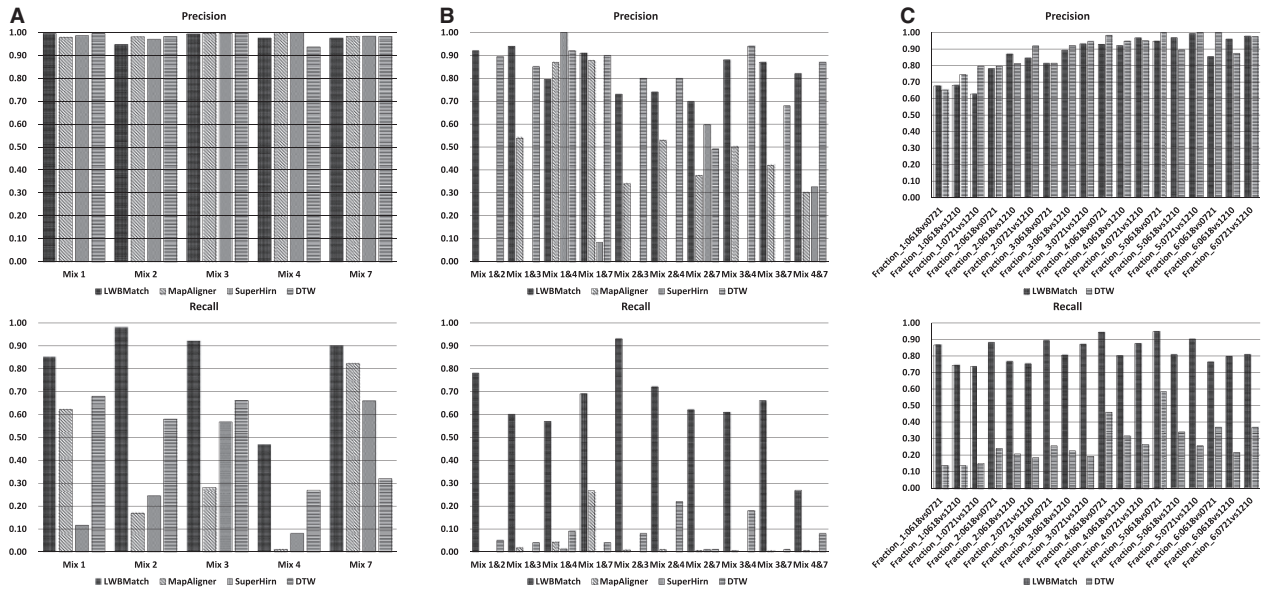

Fig. 3. (**A**) Performance of homogeneous alignments on the Standard Protein Mix datasets. (**B**) Performance of heterogeneous alignments on the Standard Protein Mix datasets. (**C**) Performance of heterogeneous alignments on U2OS datasets

indicating that LWBMatch can retrieve a higher fraction of the ground-truth pairs, without sacrificing the accuracy.

To confirm that the DTW algorithm we implemented is functioning properly, we verified that the detected optimal path fits the ground-truths well in one pair of Mix 1 datasets in Supplementary Figure S9.

4.3 Results for heterogeneous alignment

The results of heterogeneous alignment methods are shown in Figure 3B and C. Four methods are compared again in this evaluation, but because most of the recall values for MapAligner and SuperHirn are close to 0, we only list the results for LWBMatch and modified DTW. From Figure 3B and C, we observe that LWBMatch can achieve a satisfactory precision and reasonable recall, although the results are poorer compared with the homogeneous cases.

On inspection, we found that the failure of MapAligner and SuperHirn can be traced to the same problem—an inability to extract correct landmark feature pairs. MapAligner searches for landmark feature pairs within a small m/z window. In case of multiple feature pairs at the same m/z, it resolves the ambiguity by choosing the pair with the smallest RT shift. This turns out to be a poor assumption for heterogeneous alignments. In SuperHirn's case, too many incorrect landmark feature pairs are admitted when the pre-warping RT tolerance is disabled to accommodate heterogeneous alignment with large RT shifts. It has no mechanism to filter out incorrect landmark feature pairs. Given these wrong inputs, the fitted warping function (in MapAligner) or the LOWESS model (in SuperHirn) was simply incorrect.

5 DISCUSSION AND CONCLUSION

The alignment of LC-MS maps has been a well-studied topic in proteomics and a critical step in label-free quantitative proteomics, a work-flow commonly used in traditional biomarker discovery. In such experiments, one would attempt to analyze the two samples as similarly as possible, usually on the same machine and consecutively in time. As proteomics technology advances, however, the ambition of scientists and hence the size of experiments have also grown over time. Given the larger number of samples, it becomes necessary to spread the runs out over weeks or months, or use multiple instruments in parallel. In both cases, existing LC-MS alignment tools often fail, as they fail to anticipate large RT shifts or elution order reversals that can occur when the experiments are conducted in slightly different conditions. Partly due to this difficulty in LC-MS alignment, other quantitative approaches, such as stable isotope labeling methods and MS2-based spectral counting, have emerged in recent years. However, the label-free, LC-MS-based method still retains important advantages. It can be used to compare theoretically infinite number of samples, is generally applicable to all biological systems, requires no expensive stable isotope labeling reagents and is a true unbiased profiling technique. Compared with spectral counting, it allows the quantification of analytes even when MS/MS identifications were not available or not consistently present for that analyte in all experimental runs (Mueller *et al*., 2008).

Conventional LC-MS time alignment methods all focus on finding a warping function between two maps. To limit the search space for this warping function, one or more of the following assumptions are often made: (i) the warping function is monotonic, that is, peptides elute in the same order in time; (ii) matching features are within a small RT tolerance, typically several minutes; and (iii) the warping function is linear or at least locally linear. In practice, however, these assumptions are frequently violated, especially when the maps to be aligned are acquired on different chromatographic columns or under different settings. The resulting alignment is prone to error regardless of the numerical and optimization methods used to find the warping function.

To solve this problem, we developed a two-step method that approaches the challenge differently. For sufficiently similar maps, which we termed homogeneous alignments, we bypassed the step of finding the warping function entirely, and instead attempted to identify matching features directly. We note that the warping function itself is only a means to the end, which is the set of matching features across many LC-MS maps, from which biological information can be obtained. The alignment problem is instead formulated as a familiar problem in graph theory with an efficient solution. We showed that this method, WBMatch, outperforms existing feature-based methods, particularly in recall. The increased performance is likely owing to the fact that our method does not assume monotonicity or linearity of the warping function, and allows each feature to be matched independently.

To handle the more challenging 'heterogeneous' alignment tasks, such as when the LC-MS maps are acquired on different instruments, we also proposed an enhancement to WBMatch called LWBMatch. To correct large RT shifts between different maps, we applied a voting strategy to detect potential correct shifts and then LOWESS is used to fit a warping function. Although LOWESS is also used as a curve fitting strategy in some existing tools, LWBMatch allows a much wider RT tolerance and uses the co-occurrence of features at the same time as a way to minimize false matches. The warping function is then applied to one of the maps, such that the RT shifts are small enough to be aligned by WBMatch. Unlike existing tools, our method did not need to determine the warping function to a fine timescale, which was because of the effectiveness of the WBMatch step. Overall, this two-step approach assumes that elution order is only approximately conserved on a long time-scale, but can be reversed frequently within a short time-scale. We believe this assumption is closer to the reality and therefore makes our method more successful than the existing approaches.

In terms of limitation, LWBMatch currently operates on a list of features detected from LC-MS maps. Therefore, it relies on an effective algorithm for feature finding, which typically is designed for high-resolution data. Thus, at present, LWBMatch is only applicable to high-resolution datasets. Of course, for successful alignment, there must still exist sufficient similarity between the maps to be aligned. For instance, the chromatography must be based on similar chemistry. Given that nowadays almost all LC-MS proteomics experiments use reverse-phase chromatography and similar solvents, we expect our method to be applicable to most cases. However, perhaps unsurprisingly, the degree of success (as measured by precision and recall) varies from dataset to dataset and is somewhat correlated with the similarity between the maps (as measured by RT shifts and elution order reversals). Future work will focus on overcoming these limitations, and also on improving the algorithm further for the more challenging cases.

As a powerful analytical platform, LC-MS can also potentially find applications beyond traditional biomarker discovery experiments. A good LC-MS alignment algorithm can make it possible to use LC-MS maps as 'fingerprints' of biological samples, for instance to identify microorganisms by their proteome or metabolome. Finally, given that many LC-MS experiments are conducted by proteomics facilities everywhere on similar samples, often using comparable but not identical platforms, we reason that one can potentially use our method to copy peptide identifications from one map to another, thereby improving identification rates and reducing the expensive computation needed for MS2 identification. This capability can be a useful addition to proteomics data repositories.

Funding: We acknowledge the funding support of the University Grant Council of the Hong Kong Special Administrative Region Government, China (Grant No. HKUST RPC10EG08).

Conflict of Interest: none declared.

REFERENCES

Åberg,K. *et al.* (2009) The correspondence problem for metabonomics datasets. *Anal. Bioanal. Chem.*, **394**, 151–162.

America,A. and Cordewener,J. (2008) Comparative LC-MS: a landscape of peaks and valleys. *Proteomics*, **8**, 731–749.

Bisle,B. *et al.* (2006) Quantitative profiling of the membrane proteome in a halophilic archaeon. *Mol. Cell. Proteomics*, **5**, 1543–1558.

Böttcher,C. *et al.* (2008) Metabolome analysis of biosynthetic mutants reveals a diversity of metabolic changes and allows identification of a large number of new compounds in arabidopsis. *Plant Physiol.*, **147**, 2107–2120.

Catchpole,G. *et al.* (2005) Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl Acad. Sci. USA*, **102**, 14458–14462.

Christin,C. *et al.* (2008) Optimized time alignment algorithm for lc- ms data: correlation optimized warping using component detection algorithm-selected mass chromatograms. *Anal. Chem.*, **80**, 7012–7021.

Christin,C. *et al.* (2010) Time alignment algorithms based on selected mass traces for complex lc-ms data. *J. Proteome Res.*, **9**, 1483–1495.

Cleveland,W. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Dowsey,A. *et al.* (2010) Image analysis tools and emerging algorithms for expression proteomics. *Proteomics*, **10**, 4226–4257.

Eilers,P. (2004) Parametric time warping. *Anal. Chem.*, **76**, 404–411.

Eng,J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.

Finney,G.L. *et al.* (2008) Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution µlc-ms data. *Anal. Chem.*, **80**, 961–971.

Fredman,M.L. and Tarjan,R.E. (1987) Fibonacci heaps and their uses in improved network optimization algorithms. *J. ACM*, **34**, 596–615.

Geiger,T. *et al.* (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics*, **11**, M111.014050.

Hoekman,B. *et al.* (2012) msCompare: a framework for quantitative analysis of label-free LC-MS data for comparative candidate biomarker studies. *Mol. Cell. Proteomics*, **11**, M111.015974.

Jaitly,N. *et al.* (2006) Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.*, **78**, 7397–7409.

Katajamaa,M. and Orešič,M. (2005) Processing methods for differential analysis of lc/ms profile data. *BMC Bioinformatics*, **6**, 179.

Katajamaa,M. and Orešič,M. (2007) Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A*, **1158**, 318–328.

Keller,A. *et al.* (2005) A uniform proteomics ms/ms analysis platform utilizing open xml file formats. *Mol. Syst. Biol.*, **1,** 2005.0017.

Klimek,J. *et al.* (2007) The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.*, **7**, 96–103.

Kohlbacher,O. *et al.* (2007) Topp–the OpenMS proteomics pipeline. *Bioinformatics*, **23**, e191–e197.

Kuhn,H. (1955) The hungarian method for the assignment problem. *Nav. Res. Logist. Q.*, **2**, 83–97.

Lange,E. *et al.* (2007) A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics*, **23**, i273–i281.

Lange,E. *et al*. (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, **9**, 375.

Li,X. *et al*. (2005) A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics*, **4**, 1328–1340.

Listgarten,J. *et al*. (2005) Multiple alignment of continuous time series. In: *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA.

Listgarten,J. *et al*. (2007) Difference detection in LC-MS data for protein biomarker discovery. *Bioinformatics*, **23**, e198–e204.

May,D. *et al*. (2007) A platform for accurate mass and time analyses of mass spectrometry data. *J. Proteome Res*., **6**, 2685–2694.

Mueller,L. *et al*. (2007) Superhirn–a novel tool for high resolution lc-ms-based peptide/protein profiling. *Proteomics*, **7**, 3470–3480.

Mueller,L.N. *et al*. (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res*., **7**, 51–61.

Munkres,J. (1957) Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math*., **5**, 32–38.

Nielsen,S. *et al*. (2002) Triply charged bradykinin and gramicidin radical cations: their formation and the selective enhancement of charge-directed cleavage processes. *Int. J. Mass Spectrom*., **213**, 225–235.

Noy,K. *et al*. (2011) Shape-based feature matching improves protein identification via LC-MS and tandem MS. *J. Comput. Biol*., **18**, 547–557.

Prakash,A. *et al*. (2006) Signal maps for mass spectrometry-based comparative proteomics. *Mol. Cell. Proteomics*, **5**, 423–432.

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.

Sakoe,H. and Chiba,S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech, Signal Process*., **26**, 43–49.

Shteynberg,D. *et al*. (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics*, **10**, M111.007690.

Smith,C. *et al*. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem*., **78**, 779–787.

Snyder,L. and Dolan,J. (2006) *High-performance gradient elution*. Wiley-Interscience, New York, USA.

Sturm,M. *et al*. (2008) Openms–an open-source software framework for mass spectrometry. *BMC Bioinformatics*, **9**, 163.

Suits,F. *et al*. (2008) Two-dimensional method for time aligning liquid chromatography-mass spectrometry data. *Anal. Chem*., **80**, 3095–3104.

Vandenbogaert,M. *et al*. (2008) Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics*, **8**, 650–672.

Vissers,J. *et al*. (2007) Analysis and quantification of diagnostic serum markers and protein signatures for gaucher disease. *Mol. Cell. Proteomics*, **6**, 755–766.

Wang,W. *et al*. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem*., **75**, 4818–4826.

West,D.B. (2001) *Introduction to graph theory*. Vol. 2, Prentice Hall Englewood Cliffs, NJ, USA.

Zhang,X. *et al*. (2005) Data pre-processing in liquid chromatography–mass spectrometry-based proteomics. *Bioinformatics*, **21**, 4054–4059.

Zhang,Z. (2012) Retention time alignment of LC/MS data by a divide-and-conquer algorithm. *J. Am. Soc. Mass Spectrom*., **23**, 764–772.