ChimST: An Efficient Spectral Library Search Tool for Peptide Identification from Chimeric Spectra in Data-Dependent Acquisition

Wenju Zhang, Zhewei Liang, Xin Chen, Lei Xin, Baozhen Shan, Zhigang Luo, and Ming Li

Abstract—Accurate and sensitive identification of peptides from MS/MS spectra is a very challenging problem in computational shotgun proteomics. To tackle this problem, spectral library search has been one of the competitive solutions. However, most existing library search tools were developed on the basis of one peptide per spectrum, which prevents them from working properly on chimeric spectra where two or more peptides are co-fragmented. In this work, we present a new library search tool called ChimST, which is particularly capable of reliably identifying multiple peptides from a chimeric spectrum. It starts with associating each query MS/MS spectrum with MS precursor features. For each precursor feature, there is a list of peptide candidates extracted from an input spectral library. Then, it takes one peptide candidate from each associated feature and scores how well they could collectively interpret the query spectrum. The highest-scoring set of peptide candidates are finally reported as the identification of the query spectrum. Our experimental tests show that ChimST could significantly outperform the three state-of-the-art library search tools, SpectraST, reSpect, and MSPLIT, in terms of the numbers of both peptide-spectrum matches and unique peptides, especially when the acquisition isolation window is broad.

Index Terms—MS/MS, spectral library search, chimeric spectra, data-dependent acquisition

1 Introduction

Proteomics aims mainly to gain a comprehensive understanding of biological phenomena by analyzing the proteins involved in cellular processes or disease formation [1]. The high-performance liquid chromatography coupled with mass spectrometry has been the method of choice to separate and identify target protein molecules from mixture protein samples because of its high speed, sensitivity, resolving power, and dynamic range. The success of high-throughput proteomics largely relies on a computational solution to identify peptides from tandem mass spectra (MS/MS) with high accuracy and sensitivity [2]. Existing computational solutions fall into three major categories: *de novo* sequencing, which reconstructs peptide sequences without restrictions on their contents [3], [4], [5], [6], [7], [8]; database search, which scores each spectrum against the peptide sequences that occur only in a given protein database [9], [10], [11], [12], [13]; and library search, which scores each spectrum against spectra whose peptide sequences have been identified before [14], [15], [16]. Among these three solutions, spectral library search has been the most commonly adopted one because of its superior time efficiency and sensitivity [17], [18]. When applying to MS data from data-dependent acquisition (DDA) and data-independent acquisition (DIA), different algorithmic techniques are generally employed [19], [20], [21]. Our present study focuses on peptide identification on DDA spectra data, so the following discussions are limited in the scope of DDA-based library search approaches unless specifically stated otherwise.

Most library search algorithms follow the paradigm of one-spectrum-one-peptide. However, it has been well known that many acquired spectra may be chimeric, i.e., a single spectrum is actually a mixture of fragment ions from two or more co-fragmented peptides. As shown in recent studies, roughly about 50 percent of spectra may contain more than one peptide [22], [23]. Even with a narrow isolation window, there are still up to 39 percent chimeric spectra [24]. While quite a few database search tools were developed to handle chimeric spectra in the last decade [24], [25], [26], [27], [28], [29], no much research attention has so far been paid on this matter via library search. To our best knowledge, MSPLIT [30] is the only such library search tool. However, it limits itself to identifying a maximum of two peptides from a chimeric spectra and hence would unavoidably miss all the valuable identifications beyond the first two peptides. On the other hand, the Trans-Proteomic

- *W. Zhang is with the College of Computer, National University of Defense Technology, Changsha 410073, China, and also with the David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. E-mail: zhangwenju13@nudt.edu.cn.*
- *Z. Liang, X. Chen, L. Xin, and B. Shan are with Bioinformatics Solutions Inc., Waterloo, Ontario N2L 6J2, Canada. E-mail: zheweiliang@gmail.com, {xchen, lxin, bshan}@bioinfor.com.*
- *Z. Luo is with the College of Computer, National University of Defense Technology, Changsha 410073, China. E-mail: zgluo@nudt.edu.cn.*
- *M. Li is with the David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. E-mail: mli@uwaterloo.ca.*

Manuscript received 17 May 2019; revised 18 Sept. 2019; accepted 1 Oct. 2019. Date of publication 7 Oct. 2019; date of current version 6 Aug. 2021. (Corresponding authors: Zhigang Luo and Ming Li.)
Digital Object Identifier no. 10.1109/TCBB.2019.2945954

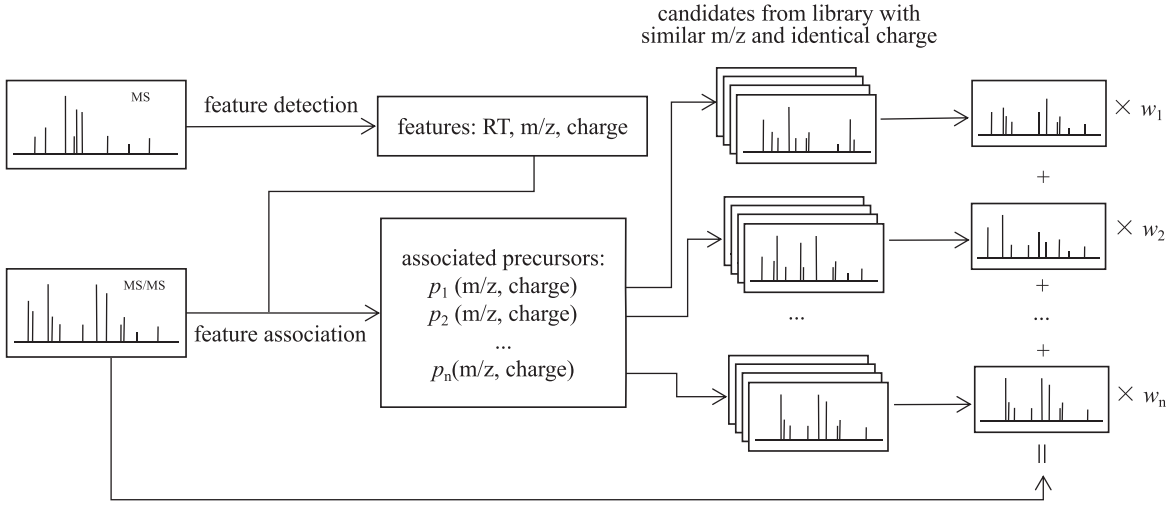


Fig. 1. The workflow of ChimST for library search from chimeric spectra. Precursor features are detected from MS spectra, with the information of retention time (RT) range, m/z, and charge. A query MS/MS spectrum is associated with a precursor feature if both m/z and RT overlap each other. For each associated precursor, the candidate spectra from a spectral library are those with matching m/z values (up to an error tolerance) and charges. Finally, peptide identification is determined from the optimal set of candidate library spectra whose non-negative weighted sum approximates the query spectrum best.

Pipeline (TPP) [31] provides a post-search processing tool called reSpect [26], which enables an iterative analysis of a chimera spectrum when coupled with a general database (or spectral library) search engine like SEQUEST [9].

To reliably and efficiently identify multiple peptides from chimeric spectra, we implemented a new library search computational approach in a tool called ChimST. Conceptually, ChimST assumes that every chimera spectrum has come from a non-negative linear combination of multiple single-peptide spectra [19], [30]. Briefly, ChimST first associates each query MS/MS spectrum with all possible precursor features detected from MS spectra and extracts for each associated feature a list of peptide/spectrum candidates from an input special library. Then, by the one-peptide-per-feature principle, it takes one peptide/spectrum candidate from each feature and scores how well they could collectively interpret the query spectrum via a convex optimization formulation. The matching score s between the query spectrum and peptide/spectrum candidates is basically calculated from the residue ϵ of a convex optimization solution (i.e., $s = 1 - \epsilon$). Various algorithmic techniques have been employed to speed up the matching process, including top-k filtering and branch-and-bound.

2 Methods

2.1 Workflow of ChimST

Fig. 1 shows our ChimST workflow. First, precursor features are detected from the survey MS spectra and their m/z, charge and retention time ranges are recorded. Then, we associate each query MS/MS spectrum to as many precursor features as possible. A successful association is achieved if a feature's m/z falls in the query spectrum's isolation window while the spectrum's retention time falls in the feature's retention time range. Afterwards, for each associated precursor we extract a candidate peptide/spectrum list from the input spectral library by matching their m/z and charge values (up to a predefined mass error tolerance). To identify a query MS/MS spectrum, we evaluate all the possible collections of library spectrum candidates in which each spectrum comes from one different associated feature, and then find out the one that would achieve the best approximation to the query spectrum via non-negative linear combination. After trimming off those significantly low weighted spectra, the remaining library spectra in the collection of best approximation will finally have their peptide identifications transferred to the query spectrum. The details of our score function and search algorithm are elucidated in Section 2.4.

2.2 Consensus Spectral Library Construction and Software Implementation

Our consensus spectral library is basically created in the same way as reported in the previous paper [32], with several quality filters being applied on the library prior to the search process. ChimST was fully written in Python and implemented various functional modules including spectral pre-processing, consensus library creation, and library search from chimeric spectra. It makes use of several open-source Python libraries such as Pyteomics, NumPy, SciPy, and pandas. The feature detection and association is done with the proteomics analysis software suite PEAKS Studio (http://bioinfor.com/). All our Python code is released as open source available at https://github.com/wj-zhang/spectral-library.

2.3 Spectrum Preprocessing

Query and library spectra are essentially preprocessed in a same way. First, we remove peaks whose mass values are smaller than 150 Da or larger than the precursor's total mass [33], as well as peaks with the relative intensity of less than 1 percent. Then, discard all spectra that meet any of the following three conditions: 1) a spectrum that contains less than 10 peaks, 2) a library spectrum for which the number of peaks is less than the length of its identified peptide, and 3) a spectrum whose mass range is less than

250 Da. Finally, peak intensities are rank transformed to de-emphasize overly dominant peaks and further normalized by the magnitude of the spectral vector [14].

2.4 Similarity Scoring

As shown in Fig. 1, a query chimera spectrum may be assigned with multiple lists of candidate library spectra through its associated precursor features. We then aim to find out library spectra from these candidates that truly carry the peptide identification of the query spectrum. Suppose that a chimeric spectrum has n precursor features and hence n candidate library spectrum lists. We consider every feasible combination of n library spectra by taking one spectrum from each candidate list, and find the best one by solving the following combinatorial optimization problem.

$$score = 1 - \min_{H} \min_{w \geq 0} \|M - Hw\|_2^2 \tag{1}$$

where M is the query spectrum vector, H is the matrix constructed by stacking the candidate library spectrum vectors, and w is a weight vector of length equal to the number of candidate library spectra under evaluation. The objective function value will be used in ChimST as a similarity score that measures the goodness of the library spectrum combination to approximate the query spectrum. Here, we use a dynamic programming-based alignment approach, rather than a common binning approach [14], to construct a unit vector representation for each spectrum. Also note that the least squares sub-problem $\min_{w \geq 0} \|M - Hw\|_2^2$ of (1) is convex because $H^T H$ is positive semi-definite and the constraint $w \geq 0$ gives a convex feasible set [34]. More details of the problem formulation are given in Supplementary Section A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2019.2945954.

Apparently, to solve the above optimization problem over all the possible library spectrum combinations, the search space increases exponentially with the increasing number of the associated precursor features. Following [30], we apply the top-k filter on candidate library spectrum lists, together with the technique of branch-and-bound, to reduce the search space and accelerate the searching process. As the pruning is based on the theoretical upper bound of all the feasible solutions, the optimal solution is guaranteed not to be trimmed off. In ChimST, we used a Python optimization toolkit called CVXOPT (https://cvxopt.org) to assist in solving the above convex optimization sub-problem. More details about the optimization algorithm are provided in Supplementary Section B, available online.

The weights in the linear combination can be naturally interpreted as the quantitative contribution of each library spectrum to the query chimera spectrum. Hence, a library spectrum with a weight value close to zero shall be considered spurious rather than a true identification. To alleviate such a noise impact, we discard all the low-weighted library spectra in the optimal candidate combination. All weights are first normalized by the maximum weight among them, and a library spectrum is then discarded if its resulting relative weight is less than a predefined threshold value of 0.15. The remaining spectra in the optimal candidate combination are finally used to determine the peptide identification of the query spectrum, i.e., simply transfer the peptide identifications of those library spectra to the query spectrum. We notice other approaches to making use of these learned weights in the literature. MSPLIT used them to train two support vector machine (SVM) models in order to re-score the two library spectrum candidates of a query spectrum, while Specter instead viewed the weights of a library spectrum across all the query spectra as a time series for subsequent linear discriminant analysis. In contrast, simply trimming off low-weighted spectra based on a reasonable threshold value is a straightforward and effective way to remove spurious identifications, as our experiments later show.

2.5 Control of False Discovery Rate

False discovery rate (FDR) is defined as the proportion of reported positive identifications that are false. In ChimST, we adopt the target-decoy strategy to estimate FDR, as commonly did in many other library search tools [14]. In brief, one decoy spectrum is first generated for each target spectrum using the shuffle-and-reposition method [35]. Then, for each query spectrum, we search the target spectrum library and decoy spectrum library separately for their respective optimal candidate spectrum combinations as described in the preceding section. The two resulting candidate spectrum combinations are further compared, and the one with a higher score is used to determine the peptide identification of the query spectrum. The identification of a query spectrum is deemed false if its candidate spectrum combination comes from the decoy library and true if otherwise. Finally, all the identified spectra are sorted according to their search scores (i.e., those given by Equation (1)) to establish FDR. Note that this target-decoy strategy could effectively avoid mixing target and decoy spectra in the identification of a query spectrum. Also, it would collapse to the traditional one when every query spectrum has only one precursor.

3 Experiments

To evaluate the performance of ChimST, we compared it with SpectraST, reSpect, and MSPLIT on two public mass spec experimental datasets and one simulated dataset.

3.1 Datasets and Library Construction

HeLa. The first dataset was generated from HeLa samples and released by [24]. Different isolation widths (2 m/z, 4 m/z, and 8 m/z) and gradient times (1 h and 3 h) were employed during data acquisition. By pairing the isolation width and gradient time, we chose to experiment on 6 sets of data, i.e., 2 m/z and 1 h, 2 m/z and 3 h, 4 m/z and 1 h, 4 m/z and 3 h, 8 m/z and 1 h, and 8 m/z and 3 h. Each data set contains three replicates. The raw data was downloaded from the PRIDE [36] partner data repository (with the identifier PXD007750). More details about these datasets can be found in [24].

In our experiments, the dataset of the 2 m/z isolation window and 3 h gradient time was chosen to build the library, while the remaining five were used for the library searching process. To build a spectral library, we ran PEAKS DB [13] to obtain peptide identifications. The database search parameters were set as suggested by [24].

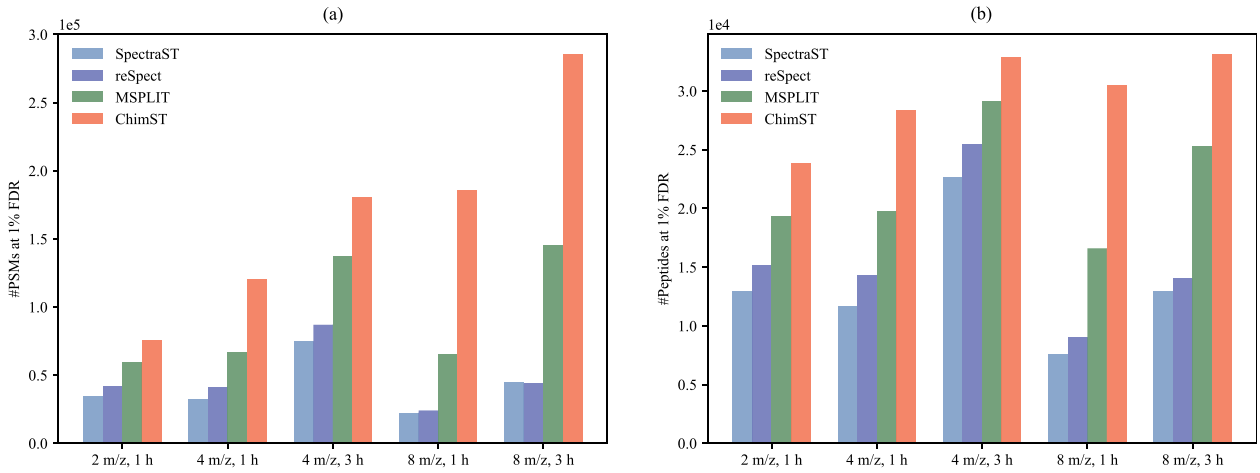


Fig. 2. Comparison of identification results of SpectraST, reSpect, MSPLIT, and ChimST on the HeLa dataset under various isolation widths and gradient times. Numbers of PSMs identified at 1 percent FDR are shown on the left and numbers of unique peptides on the right.

Specifically, the mass error tolerance was 10 ppm for precursor ions and 0.02 Da for product ions. Trypsin enzyme and 2 maximal missed cleavages were specified. Carbamidomethyl (C) was used as a fixed modification and Oxidation (M) as a variable modification. High-quality peptide-spectrum matches (PSMs) were identified at 0.1 percent FDR and then passed to both ChimST and SpectraST to build library, resulting respectively in 38,838 and 34,063 entries (i.e., pairs of peptide sequence and charge). MSPLIT will use the library built by SpectraST because it does not have its own library building module.

iPRG2017. The second dataset was released by the Proteomics Research Group of the Association of Biomolecular Resource Facilities in 2017. The raw data contains 4 LC-MS/MS files. We built the spectral library on the first two data files, using the same parameter settings as those for the HeLa dataset except that Deamidation (NQ) is also included as a variable modification. ChimST generated 29,378 entries in the library and SpectraST generated 26,947 entries. Library search was then conducted on the other two data files in our experiments below.

Simulated dataset. Because there are no real chimeric spectra data with their verified peptide identifications publicly available, we created a simulated chimeric spectra dataset to test the applicability of our proposed library search method. To do so, we used spectra in the spectral library built from HeLa dataset as the base single-peptide spectra, and then generated each simulated chimeric spectrum by mixing one up to five selected base spectra with different weights. The number of selected base spectra defines the number of co-fragmented peptides and hence the number of precursor features associated with the simulated chimeric spectrum. Set an isolation window of ± 2.5 m/z, so all the precursor masses of a chimeric spectrum are restricted in the range of ± 2.5 m/z. Since our method relies on the quality of associated precursors of query spectra, we further simulated noises by adding three spurious spectra/precursors to each chimeric spectrum with a low mixing weight of 0.05. Any identification of such a spurious spectrum/precursor will be considered as a false positive. However, these spurious spectra/precursors will be removed from the chimeric spectra when we test other methods (i.e., SpectraST, reSpect, and MSPLIT), making them immune to noises. For detailed information about the simulated dataset, please see Supplementary Table S2, available online.

3.2 Parameter Settings

For both real and simulated spectral libraries, the decoy spectra were generated at the ratio of 1:1 to target spectra for FDR estimate. For the library search by ChimST, both the precursor ion mass tolerance and the product ion m/z tolerance were set to the default value of 25 ppm for all datasets. To run SpectraST and MSPLIT, the precursor ion mass tolerances were set to the isolation width for the HeLa dataset and the default value of 3.0 Da for both iPRG2017 and simulated datasets. For reSpect, SpectraST was used as the search engine, with three rounds of search for real datasets as advised in [26] and with four rounds of search for the simulated dataset because the simulated spectra have maximal 5 co-fragmented peptides. In addition, as reSpect recommended [26], a smaller precursor mass tolerance of 1.1 Da was adopted for the first round of search, and the search results were further processed with PeptideProphet and iProphet to distinguish targets and decoys. SpectraST, reSpect, PeptideProphet, and iProphet are all from the Trans-Proteomic Pipeline version 5.1.0 [31].

4 Results and Discussion

4.1 Results on the HeLa Dataset

To evaluate the performance of ChimST relative to SpectraST, reSpect, and MSPLIT on the HeLa dataset, we compared the numbers of PSMs identified at the 1 percent FDR and the numbers of unique peptides therein reported.

As demonstrated in Fig. 2, ChimST outperforms SpectraST, reSpect, and MSPLIT in terms of both the number of PSMs and the number of unique peptides at 1 percent FDR for all tested isolation widths and gradient times. Even for the narrow isolation width and small gradient time (2 m/z and 1h), ChimST identifies up to 27.7 percent more PSMs and 23.4 percent more unique peptides than the best baseline MSPLIT. ChimST increases the number of PSMs significantly by 184.9 percent and unique peptides by 83.7 percent when compared with MSPLIT for a broad isolation width (8 m/z and 1 h).

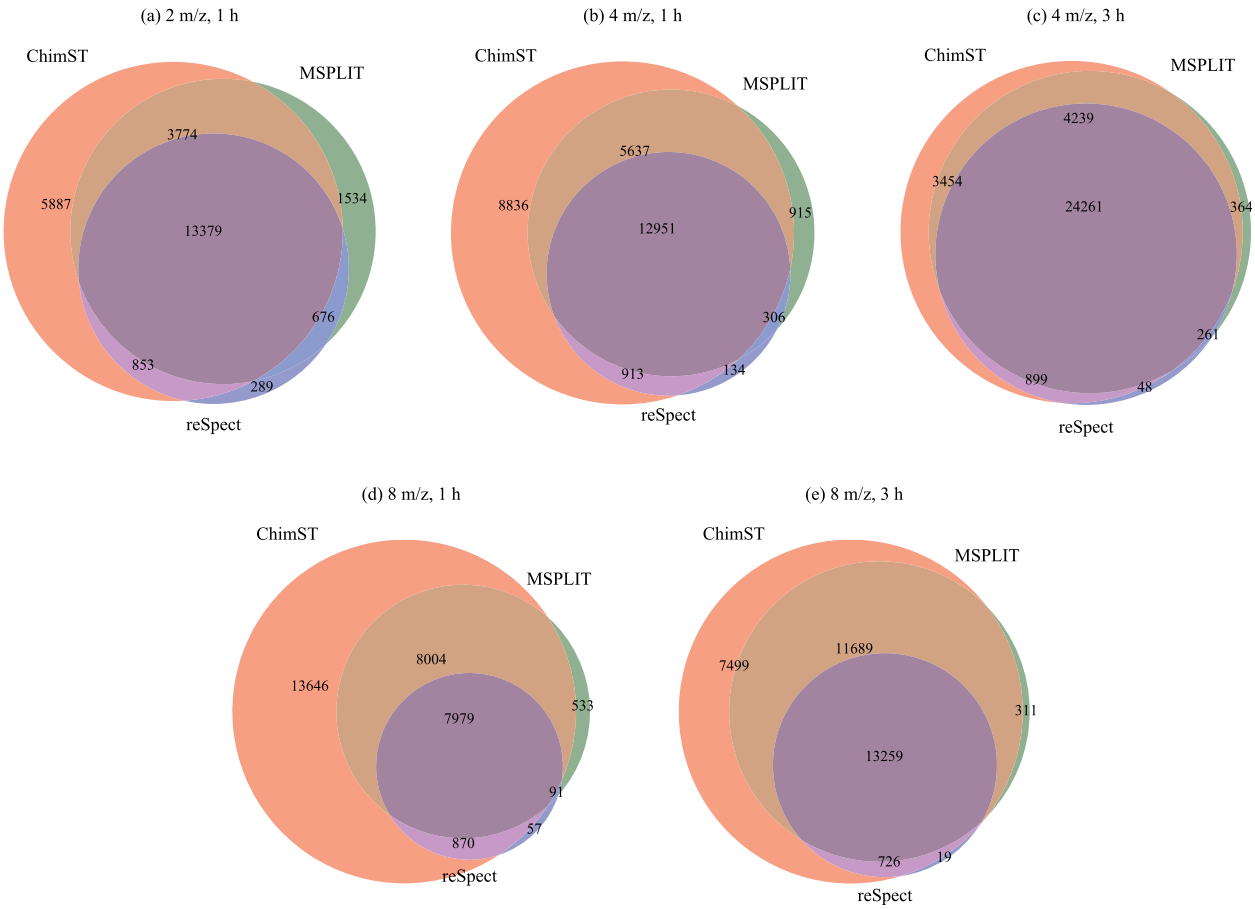


Fig. 3. Venn diagrams of the numbers of unique peptides at 1 percent FDR from reSpect, MSPLIT, and ChimST on the HeLa dataset under different isolation widths and gradient times.

Generally, we shall expect more PSMs and peptides to be identified when the isolation width increases, because a wider isolation window enables more peptides co-fragmented. However, SpectraST exhibits the opposite trend in that the amount of identified PSMs and peptides are actually decreasing when the isolation width is increasing. This is not surprising, because SpectraST was designed to identify one peptide only from each spectrum. It will definitely miss the identification of all but one peptide from a chimeric spectrum. Moreover, the presence of fragment ions from other peptides may make SpectraST hard to identify any confident peptide identification owing to its PSM scoring function.

For the case of MSPLIT, it displays a hump pattern: the number of identified PSMs and peptides increase as the isolation width grows from 2 m/z to 4 m/z but decrease when the width goes from 4 m/z to 8 m/z. Compared to SpectraST, MSPLIT was designed to deal with two-peptide chimeric spectra, which explains why performance improvement can be seen when the isolation width rises from 2 m/z to 4 m/z. The performance starts to drop after the isolation width of 4 m/z, because there are more chimeric spectra expected from three or more co-fragmented peptides.

On the other hand, reSpect can identify more peptides than SpectraST on all tested datasets (Fig. 2b), which shows its ability to resolve chimeric spectra. However, it still performs no better than MSPLIT and ChimST. We believe this is because reSpect relies on SpectraST for peptide identification and therefore can not get around the limitation of SpectraST that scores the matching between a single peptide and a mixture spectrum.

For a better illustration, Fig. 3 shows the overlapping results of identified peptides by ChimST, reSpect and MSPLIT on the HeLa dataset. The 4-set Venn diagrams of identified peptides from the four tested methods are given in Supplementary Fig. S3, available online. As we can see, almost all the reSpect's peptide identifications (from 98.1 to 99.9 percent) can be found in either MSPLIT's or ChimST's results; most of MSPLIT's identifications (from 88.6 to 98.8 percent) can be found in ChimST's results; and ChimST is able to identify more unique peptides than MSPLIT (from 14.9 to 87.4 percent). In particular, as the isolation width increases, both reSpect and MSPLIT tend to identify less unique peptides but ChimST continues to find more, which strongly indicates that ChimST can resolve highly complex chimeric spectra very well.

Statistical analysis was performed on identified spectra from ChimST (Supplementary Fig. S5, available online), showing that 36 percent of identified spectra from the data set of 2 m/z isolation width and 1 h gradient time are chimeric spectra and 9 percent contain more than two co-fragmented peptides. These two numbers arise to 73 and 44 percent respectively at the broad isolation width 8 m/z and 3 h. Therefore, it is very demanding to develop a new spectral library search tool like ChimST that can efficiently tackle chimeric spectra.

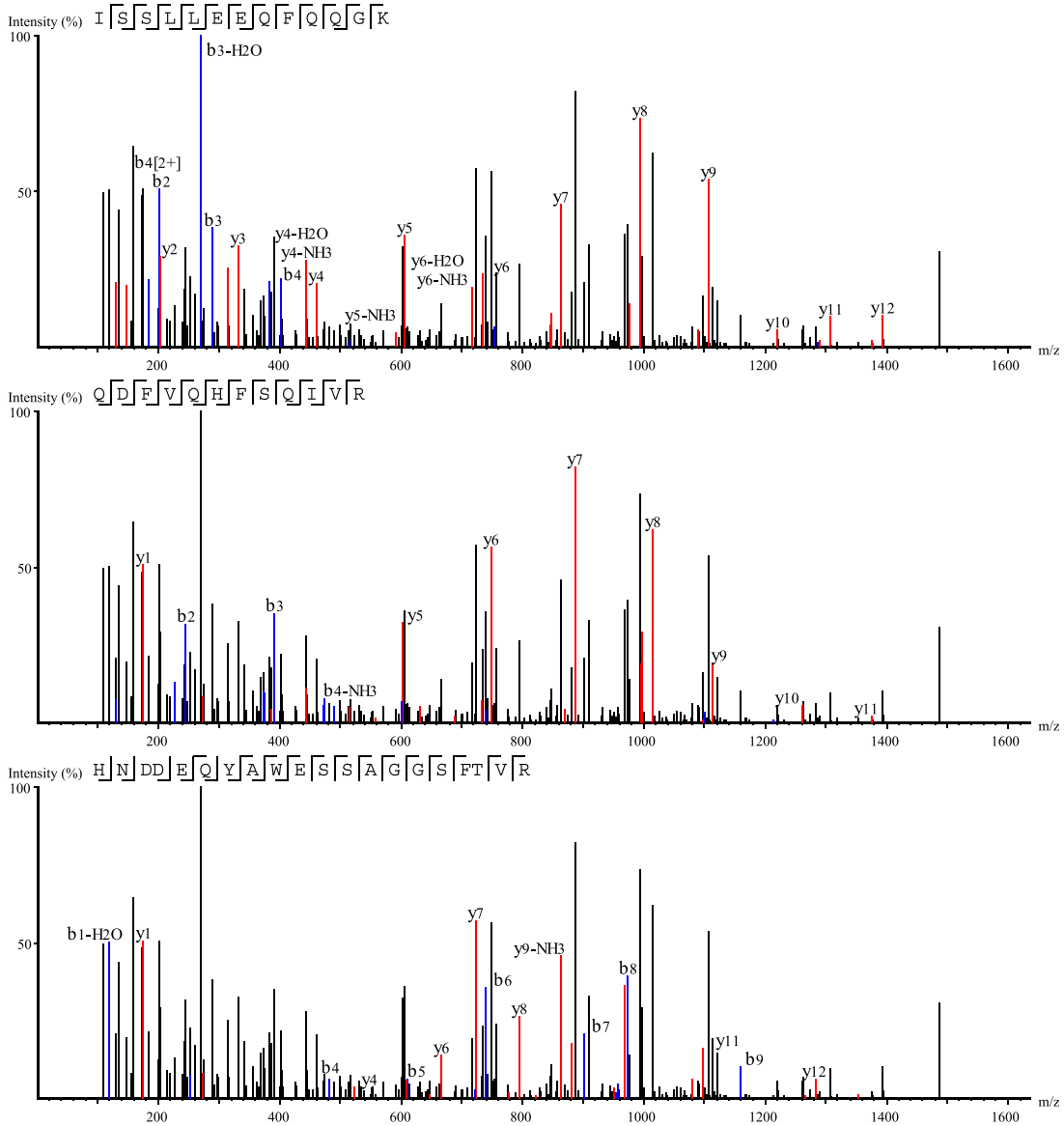


Fig. 4. Example of a chimeric spectrum (scan number of 19574 in the first replicate of data set of 8 m/z and 1 h). ChimST identified three co-fragmented peptides ISSLLEEQFQQGK, QDFVQHFSQIVR, and HNDDEQYAWESSAGGSFTVR whose weights are 0.53, 0.42, and 0.16, respectively. MSPLIT identified two peptides ISSLLEEQFQQGK and QDFVQHFSQIVR whereas SpectraST and reSpect only identified ISSLLEEQFQQGK. The theoretically matching fragment ions are highlighted in each panel, where y ions are colored in red and b ions in blue.

We further manually validated those peptide-spectrum matches reported only by ChimST. 20 selected examples are provided in Supplementary Fig. S10, available online. The presence of many theoretically matching fragment ions indicates that they are more likely to be true identifications than false. Another interesting example is a chimeric spectrum from the data set of 8 m/z and 1 h shown in Fig. 4. Different methods give different sets of peptides. ChimST was able to identify three peptides, all having a number of (mostly non-overlaping) theoretically matching fragment ions. In comparison, MSPLIT identified the first two peptides, while both SpectraST and reSpect uncovered the first one only.

Finally, we investigated how the application of top-k filter would accelerate the search for optimal peptide identifications in ChimST. Experimental results on the data set of 4 m/z and 1 h are shown in Fig. 5a. Generally, the top-k filter can speed up the search by a significant factor (more than 14,000 times faster when $k = 1$) while still maintaining a very high percentage (>93 percent) of optimal solution hits. The percentage of optimal solution hits can improve as k increases, but at the cost of longer search time. Experiments also show that the technique of branch-and-bound can further speed up the search, the extent of which depends largely on the value of k. It is not surprising that there is a trade-off between the percentage of optimal solution hits and the speedup in terms of k. By taking both into consideration, we find that $k = 3$ is a good choice which gives a high percentage of 99.4 percent and a large speedup of 652 folds from the top-k filter. Together with the technique of branch-and-bound, the total speedup can achieve up to 1,695 ($\approx 652 \times 2.6$) folds.

4.2 Results on the iPRG2017 Dataset

We next test the performances of SpectraST, reSpect, MSPLIT, and ChimST on the iPRG2017 dataset. Numbers of identified PSMs and unique peptides at 1 percent FDR are

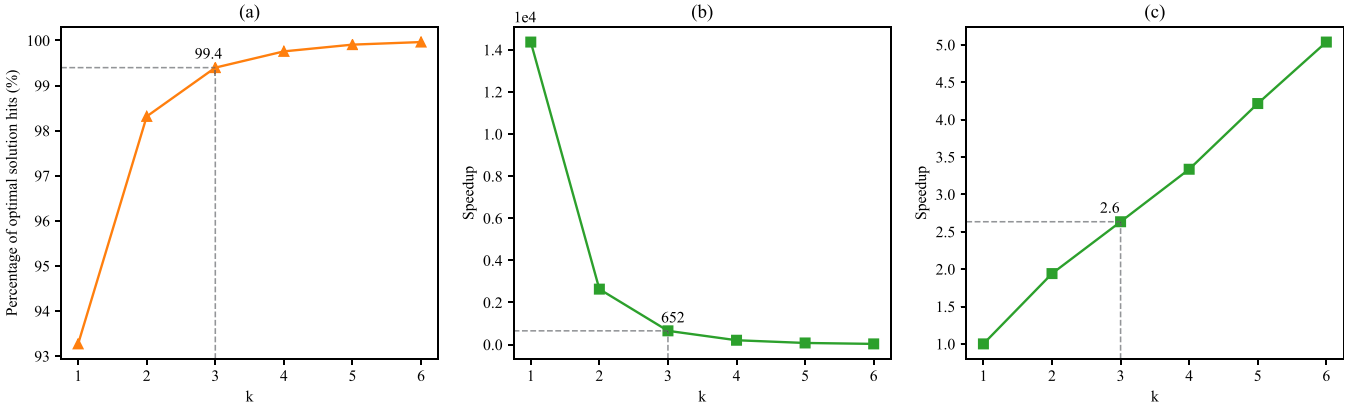


Fig. 5. Percentage of the optimal solution hits (a), speedup by top-k filter (b), and speedup by branch-and-bound after applying top-k filter (c) as a function of k when experimented on the first replicate of data set of 4 m/z and 1 h. Here, the speedup is defined as the ratio between the numbers of candidate library spectrum combinations to be evaluated during the optimization (which is directly proportional to the running time). The optimal solutions are obtained by searching the whole feasible space without applying the top-k filter.

depicted in Fig. 6. As clearly shown in the histograms, ChimST once again outperforms the other three methods by identifying more PSMs and peptides. Not surprisingly, SpectraST reported the least numbers of PSMs and unique peptides because it was designed only to identify a single peptide out of a spectrum even when the spectrum is chimeric. Although MSPLIT performs better than SpectraST and reSpect, it still reports 16.8 percent less PSMs and 15.0 percent less peptides than ChimST on average.

The Venn diagrams in Fig. 7 show how many common peptides are shared among the tested methods. As we have seen earlier, ChimST was able to identify most of the peptides (>90 percent) that reSpect and MSPLIT did. For those peptide identifications only by ChimST, our manual checks also show that they can be justified by theoretically matching fragment ions in most cases.

4.3 Results on the Simulated Dataset

Three metrics of precision, recall, and F-measure are used to evaluate the library search results:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{4}$$

where F-measure is a composite evaluation metric of the performance since it averages precision and recall.

As we can see from Fig. 8c, SpectraST achieves the highest performance (i.e., the F-measure value being almost 1) when the co-fragmented peptide number is 1. As the co-fragmented peptide number increases, its performance drops dramatically. It is mainly attributed to a large number of missing true identifications as verified by the rapidly decreasing tendency of recall. Meanwhile, the precision of SpectraST also exhibits a decreasing tendency (Fig. 8a). This is because it becomes very challenging for SpectraST to identify peptides from complex chimera spectra so that more peptides are falsely identified with the increasing co-fragmented peptide number. As expected, reSpect performs better than SpectraST when the co-fragmented peptide number is greater than 1. This confirms that reSpect is able to correctly identify additional peptides from chimeric spectra. However, reSpect still suffers from both false positives and false negatives.

From Fig. 8c, we can see that MSPLIT achieves very good performance when the co-fragmented peptide number is 1 or 2. This can be quickly justified by the fact that the simulated spectra fit well the theoretical assumption behind the

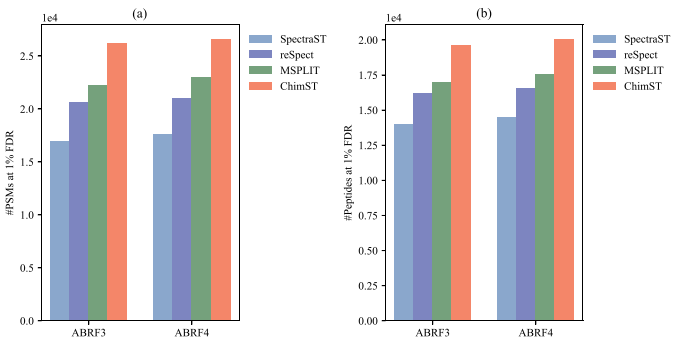


Fig. 6. Comparison of identification results of SpectraST, reSpect, MSPLIT, and ChimST on iPRG2017 datasets with various isolation widths and gradient times. Numbers of identified PSMs at 1 percent FDR are shown on the left and numbers of unique peptides on the right.

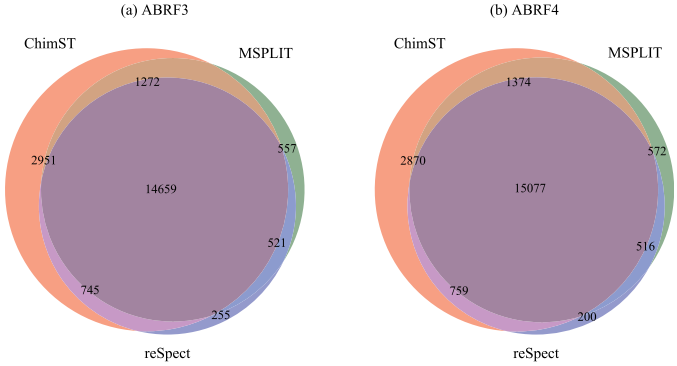


Fig. 7. Venn diagrams of the numbers of unique peptides at 1 percent FDR from reSpect, MSPLIT, and ChimST on ABRF3 and ABRF4 data sets.

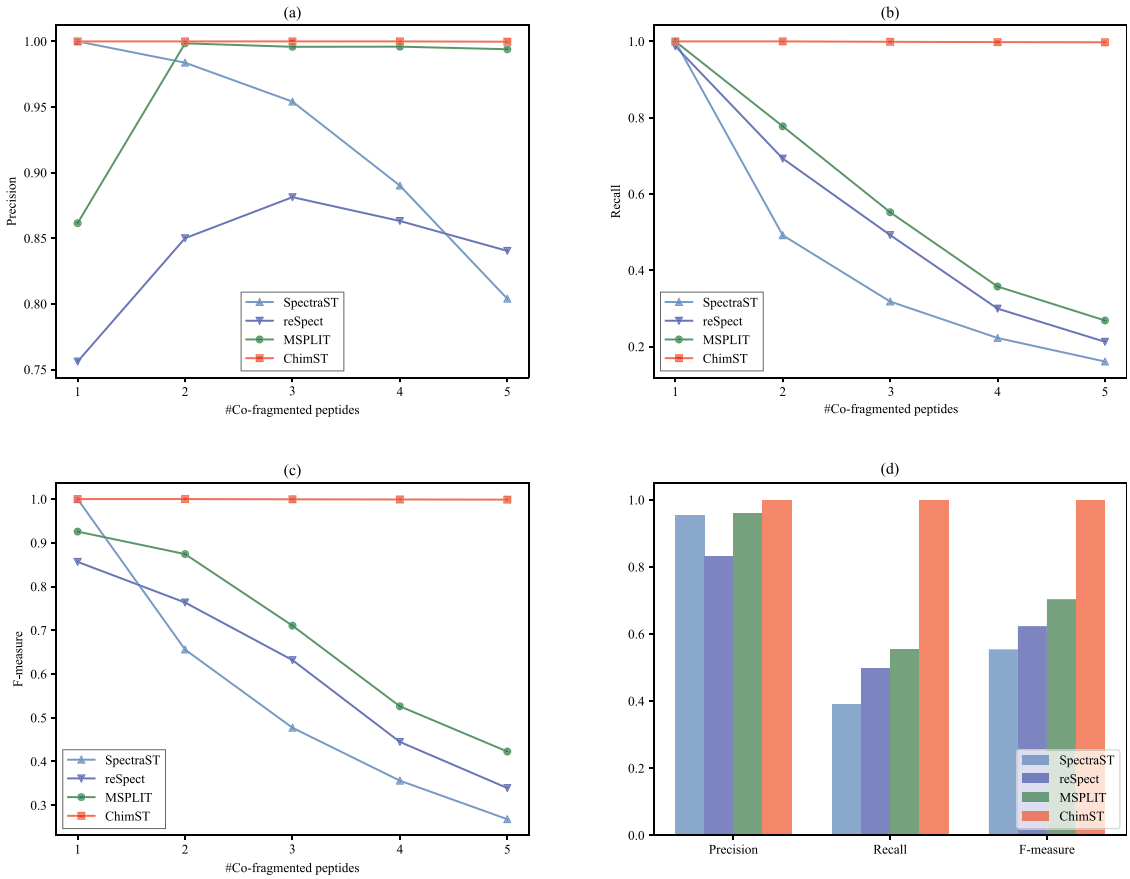


Fig. 8. Precision (a), recall (b), and F-measure (c) versus numbers of co-fragmented peptides and the overall values (d) of identification results from SpectraST, reSpect, MSPLIT, and ChimST on the simulated dataset.

algorithm of MSPLIT, i.e., a single spectrum is a mixture of one or two peptides. However, this assumption is also a major limitation of MSPLIT, which does not allow it to find three or more peptides from a single spectrum (see its low recall values shown in Fig. 8b). As a result, its performance degrades quite dramatically after the co-fragmented peptide number is 2.

In comparison, ChimST achieved the highest scores of almost 1 in all cases on the three evaluation metrics, despite the noises additionally added to mass spectra. These results clearly indicate that ChimST can not only identify true co-fragmented peptides but also learn their weights accurately. It is these accurately learned weights that enable the presumably spurious peptides to be identified and removed as we desired.

5 Conclusions

In this paper, we present a new library search method, ChimST, which is particularly capable of identifying more than two peptides from chimeric spectra in data-dependent acquisition. So far, most spectral library search methods for DDA data (e.g., SpectraST) assume that each spectrum contains only one peptide component. Consequently, they can only identify the major peptide component from a spectrum at best, even when multiple peptides may be present. reSpect and MSPLIT were instead developed to cope with chimeric spectra. However, reSpect suffers from the reliance on a library search tool still assuming one peptide per spectrum, and MSPLIT is limited to identifying at most two peptides from a single spectrum.

Different from existing methods, ChimST takes advantage of precursor features detected from MS spectra. The features associated with a spectrum provide valuable information which is beneficial to library search. First, the feature-based library spectrum candidate selection strategy can efficiently filter out unrelated spectra from library even when the isolation width is broad. Second, associated features allow to naturally set up the one-peptide-one-feature principle which restricts one possible peptide hit from each associated feature's candidate spectrum list. It would prevent many false identifications and reduce the scale of feasible candidate spectrum combinations as well. In addition, the number of associated features provides basically an upper limit on the number of peptides co-fragmented in a query spectrum. Our experiments on the two real datasets of HeLa and iPRG2017 with various isolation widths and gradient times show that ChimST largely increased the number of identified PSMs and unique peptides, especially when the isolation width is broad (184.9% more PSMs and 83.7% more peptides than MSPLIT for the data set of 8 m/z and 1 h). For the simulated dataset, ChimST could almost fully and accurately identifies all peptides from the simulated spectra whose co-fragmented peptide numbers range from 1 to 5. Overall, ChimST is an efficient library search tool for sensitive and accurate peptide identifications from chimeric spectra, and hence anticipated to be useful for gaining a deeper profiling of peptides from DDA experiments.

ACKNOWLEDGMENTS

This work was funded in part by NSERC (grant OGP0046506), China's Research and Development Program (grants 2016YFB1000902 and 2018YFB1003202), the NSFC (grant 61832019), and the Canada Research Chair Program.

REFERENCES

[1] W. P. Blackstock and M. P. Weir, "Proteomics: Quantitative and physical mapping of cellular proteins," *Trends Biotechnology*, vol. 17, no. 3, pp. 121–127, 1999.

[2] A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics," *J. Proteomics*, vol. 73, no. 11, pp. 2092–2123, 2010.

[3] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, "PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrometry*, vol. 17, no. 20, pp. 2337–2342, 2003.

[4] A. Frank and P. Pevzner, "PepNovo: de novo peptide sequencing via probabilistic network modeling," *Analytical Chemistry*, vol. 77, no. 4, pp. 964–973, 2005.

[5] B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, and J. M. Buhmann, "NovoHMM: a hidden markov model for de novo peptide sequencing," *Analytical Chemistry*, vol. 77, no. 22, pp. 7265–7273, 2005.

[6] L. Mo, D. Dutta, Y. Wan, and T. Chen, "MSNovo: A dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry," *Analytical Chemistry*, vol. 79, no. 13, pp. 4870–4878, 2007.

[7] N. H. Tran, X. Zhang, L. Xin, B. Shan, and M. Li, "De novo peptide sequencing by deep learning," *Proc. Nat. Acad. Sci.*, vol. 114, no. 31, pp. 8247–8252, 2017.

[8] Y. Liu, B. Ma, K. Zhang, and G. Lajoie, "An approach for peptide identification by de novo sequencing of mixture spectra," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 2, pp. 326–336, Mar./Apr. 2017.

[9] J. K. Eng, A. L. McCormack, and J. R. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Amer. Soc. Mass Spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.

[10] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *ELECTROPHORESIS: An Int. J.*, vol. 20, no. 18, pp. 3551–3567, 1999.

[11] R. Craig and R. C. Beavis, "TANDEM: Matching proteins with tandem mass spectra," *Bioinf.*, vol. 20, no. 9, pp. 1466–1467, 2004.

[12] S. Tanner, H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner, and V. Bafna, "InsPecT: identification of posttranslationally modified peptides from tandem mass spectra," *Analytical Chemistry*, vol. 77, no. 14, pp. 4626–4639, 2005.

[13] J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. A. Lajoie, and B. Ma, "PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification," *Mol. Cellular Proteomics*, vol. 11, no. 4, pp. M111–010 587, 2012.

[14] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, and R. Aebersold, "Development and validation of a spectral library searching method for peptide identification from MS/MS," *Proteomics*, vol. 7, no. 5, pp. 655–667, 2007.

[15] B. E. Frewen, G. E. Merrihew, C. C. Wu, W. S. Noble, and M. J. MacCoss, "Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries," *Analytical Chemistry*, vol. 78, no. 16, pp. 5678–5684, 2006.

[16] R. Craig, J. Cortens, D. Fenyo, and R. C. Beavis, "Using annotated peptide mass spectrum libraries for protein identification," *J. Proteome Res.*, vol. 5, no. 8, pp. 1843–1849, 2006.

[17] H. Lam, "Building and searching tandem mass spectral libraries for peptide identification," *Mol. Cellular Proteomics*, vol. 10, no. 12, pp. R111–008 565, 2011.

[18] X. Zhang, Y. Li, W. Shao, and H. Lam, "Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis," *Proteomics*, vol. 11, no. 6, pp. 1075–1085, 2011.

[19] R. Peckner, S. A. Myers, A. S. V. Jacome, J. D. Egertson, J. G. Abelin, M. J. MacCoss, S. A. Carr, and J. D. Jaffe, "Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics," *Nature Methods*, vol. 15, no. 5, 2018, Art. no. 371.

[20] J. Wang, M. Tucholska, J. D. Knight, J.-P. Lambert, S. Tate, B. Larsen, A.-C. Gingras, and N. Bandeira, "MSPLIT-DIA: sensitive peptide identification for data-independent acquisition," *Nature Methods*, vol. 12, no. 12, 2015, Art. no. 1106.

[21] C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras, and A. I. Nesvizhskii, "DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics," *Nature Methods*, vol. 12, no. 3, 2015, Art. no. 258.

[22] A. Michalski, J. Cox, and M. Mann, "More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS," *J. Proteome Res.*, vol. 10, no. 4, pp. 1785–1793, 2011.

[23] J. Wang, P. E. Bourne, and N. Bandeira, "MixGF: Spectral probabilities for mixture spectra from more than one peptide," *Mol. Cellular Proteomics*, vol. 13, no. 12, pp. 3688–3697, 2014.

[24] V. Dorfer, S. Maltsev, S. Winkler, and K. Mechtler, "CharmeRT: boosting peptide identifications by chimeric spectra identification and retention time prediction," *J. Proteome Res.*, vol. 17, no. 8, pp. 2581–2589, 2018.

[25] B. Zhang, M. Pirmoradian, A. Chernobrovkin, and R. A. Zubarev, "DeMix workflow for efficient identification of cofragmented peptides in high resolution data-dependent tandem mass spectrometry," *Mol. Cellular Proteomics*, vol. 13, no. 11, pp. 3211–3223, 2014.

[26] D. Shteynberg, L. Mendoza, M. R. Hoopmann, Z. Sun, F. Schmidt, E. W. Deutsch, and R. L. Moritz, "reSpect: software for identification of high and low abundance ion species in chimeric tandem mass spectra," *J. Amer. Soc. Mass Spectrometry*, vol. 26, no. 11, pp. 1837–1847, 2015.

[27] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "Andromeda: a peptide search engine integrated into the MaxQuant environment," *J. Proteome Res.*, vol. 10, no. 4, pp. 1794–1805, 2011.

[28] J. Wang, P. E. Bourne, and N. Bandeira, "Peptide identification by database search of mixture tandem mass spectra," *Mol. Cellular Proteomics*, vol. 10, no. 12, pp. M111–010 017, 2011.

[29] N. Zhang, X.-j. Li, M. Ye, S. Pan, B. Schwikowski, and R. Aebersold, "ProbIDtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer," *Proteomics*, vol. 5, no. 16, pp. 4096–4106, 2005.

[30] J. Wang, J. Pérez-Santiago, J. E. Katz, P. Mallick, and N. Bandeira, "Peptide identification from mixture tandem mass spectra," *Mol. Cellular Proteomics*, vol. 9, no. 7, pp. 1476–1485, 2010.

[31] E. W. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, H. Lam, N. Tasman, Z. Sun, E. Nilsson, B. Pratt, B. Prazen, et al., "A guided tour of the trans-proteomic pipeline," *Proteomics*, vol. 10, no. 6, pp. 1150–1159, 2010.

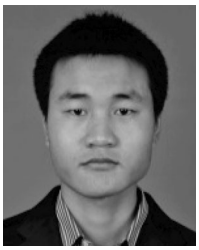
[32] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, S. E. Stein, and R. Aebersold, "Building consensus spectral libraries for peptide identification in proteomics," *Nature Methods*, vol. 5, no. 10, 2008, Art. no. 873.

[33] B. Y. Renard, M. Kirchner, F. Monigatti, A. R. Ivanov, J. Rappsilber, D. Winter, J. A. Steen, F. A. Hamprecht, and H. Steen, "When less can yield more–computational preprocessing of MS/MS spectra for peptide identification," *Proteomics*, vol. 9, no. 21, pp. 4978–4984, 2009.

[34] V. Franc, V. Hlaváč, and M. Navara, "Sequential coordinate-wise algorithm for the non-negative least squares problem," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2005, pp. 407–414.

[35] H. Lam, E. W. Deutsch, and R. Aebersold, "Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics," *J. Proteome Res.*, vol. 9, no. 1, pp. 605–610, 2009.

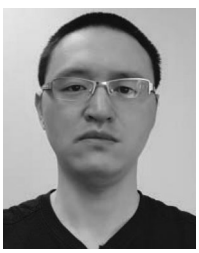
[36] J. A. Vizcaíno, A. Csordas, N. Del-Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, et al., "2016 update of the pride database and its related tools," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D447–D456, 2015.



Wenju Zhang received the BS and MS degrees in computer science and technology from the Beijing Institute of Technology, Beijing, China, and the National University of Defense Technology, Changsha, China, in 2013, and 2015, respectively. He is currently working toward the PhD degree in computer science and technology from the National University of Defense Technology, Changsha, China. He is now a visiting PhD student at the David R. Cheriton School of Computer Science, University of Waterloo. His current research interests include machine learning, transfer learning, and bioinformatics.

Zhewei Liang received the BEng degree in computer science & engineering from Xiangtan University (XTU), Hunan, China, in 2001 and the MSc and PhD degree from the Department of Computer Science, The University of Western Ontario (UWO), Ontario, Canada, in 2011 and 2017, respectively. He was a teaching assistant and an assistant professor with XTU until 2009. He worked as a postdoctoral fellow with UWO in 2018. Currently, he is a research scientist and a software developer with Bioinformatics Solutions, Inc. His research interests include bioinformatics and computational biology.

Xin Chen received the PhD degree in applied mathematics from Peking University, Beijing, China, in 2001. Currently, he is a senior research scientist with Bioinformatics Solutions, Inc., Waterloo, Ontario, Canada.



Lei Xin received the BS degree in mathematics from Peking University, in 2001, the MS degree in applied mathematics from Peking University, in 2004, and the PhD degree in computer science from the Western University, in 2010. He is now the chief technology officer with Bioinformatics Solutions, Inc. His research interests include in the field of large scale proteomics data analysis algorithm, high performance computing, and computational biology.

Baozhen Shan received the first PhD degree in chemistry from Nanjing University, in 1997, and the second PhD degree in computer science from the University of Western Ontario, in 2009. He is the CEO of BSI, one of the leading proteomics software and service platform providing companies. He began working with BSI in 2008. Currently, his team is working on the proteomics-based immunotherapy with AI technology. As a computational chemist, he has spent more than 15 years in research and development related to data mining and statistical analysis of mass spectrometry-based proteomics.

Zhigang Luo received the BS, MS, and PhD degrees from the National University of Defense Technology, Changsha, China, in 1981, 1993, and 2000, respectively. He is currently a professor with the College of Computer, National University of Defense Technology. His current research interests include machine learning, computer vision, and bioinformatics.

Ming Li received the PhD degree in computer science from Cornell University, in 1985. He is a Canada research chair in bioinformatics and a university professor with the University of Waterloo. He is a recipient of Canada's E.W.R. Steacie Fellowship Award in 1996, the 2001 Killam Fellowship, and the 2010 Killam Prize. His research interests include deep learning, conversation robots, Kolmogorov complexity and its applications, analysis of algorithms, computational complexity, natural language processing, and bioinformatics. He is a fellow of the Royal Society of Canada, ACM, and IEEE.