RESEARCH ARTICLE

pParse: A method for accurate determination of monoisotopic peaks in high-resolution mass spectra

Zuo-Fei Yuan[1,2], Chao Liu[1,2], Hai-Peng Wang[1], Rui-Xiang Sun[1], Yan Fu[1], Jing-Fen Zhang[1], Le-Heng Wang[1], Hao Chi[1,2], You Li[1], Li-Yun Xiu[1], Wen-Ping Wang[1] and Si-Min He[1]

[1] Key Laboratory of Intelligent Information Processing – Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P. R. China
[2] Graduate University of the Chinese Academy of Sciences, Beijing, P. R. China

Received: February 10, 2011
Revised: October 31, 2011
Accepted: November 2, 2011

Determining the monoisotopic peak of a precursor is a first step in interpreting mass spectra, which is basic but non-trivial. The reason is that in the isolation window of a precursor, other peaks interfere with the determination of the monoisotopic peak, leading to wrong mass-to-charge ratio or charge state. Here we propose a method, named pParse, to export the most probable monoisotopic peaks for precursors, including co-eluted precursors. We use the relationship between the position of the highest peak and the mass of the first peak to detect candidate clusters. Then, we extract three features to sort the candidate clusters: (i) the sum of the intensity, (ii) the similarity of the experimental and the theoretical isotopic distribution, and (iii) the similarity of elution profiles. We showed that the recall of pParse, MaxQuant, and BioWorks was 98–98.8%, 0.5–17%, and 1.8–36.5% at the same precision, respectively. About 50% of tandem mass spectra are triggered by multiple precursors which are difficult to identify. Then we design a new scoring function to identify the co-eluted precursors. About 26% of all identified peptides were exclusively from co-eluted peptides. Therefore, accurately determining monoisotopic peaks, including co-eluted precursors, can greatly increase peptide identification rate.

Keywords:
Bioinformatics / Co-eluted precursors / High resolution / Mass spectra / Monoisotopic peaks

1 Introduction

Peptide identification has become a key technique in MS-based proteomics [1, 2]. The main process is as follows: protein samples are proteolytically digested into peptides that are subsequently separated by liquid chromatography (LC) and then dynamically selected for fragmentation by mass spectrometers; the resultant MS/MS spectra are searched against a database to produce peptide-spectrum matches (PSM). In a database search, the monoisotopic mass and mass tolerance for a precursor can be used to obtain candidate peptides. Thanks to modern technology, precursor mass tolerance of part per million (ppm)-level can be achieved in high-resolution mass spectrometers such as the FT-ICR, Orbitrap, and orthogonal TOF instruments [3–5], which can be used to directly reduce the number of candidate peptides. In fact, dozens of candidate peptides can be obtained with several ppm, while thousands of candidate peptides will be obtained with several Daltons (Da). Unfortunately, when the given precursor mass is not the monoisotopic one, the correct peptide will not fall in the precursor mass window with ppm-level mass tolerance. Therefore, accurate determination of monoisotopic masses for precursors is important for peptide identification in high-resolution mass spectra.

Correspondence: Dr. Si-Min He, Institute of Computing Technology, Chinese Academy of Sciences, No. 6 Kexueyuan South Road, Haidian District, Beijing 100190, P. R. China
E-mail: smhe@ict.ac.cn
Fax: +86-10-62601356

Abbreviations: FDR, false discovery rate; **ppm,** part per million; **PSM,** peptide-spectrum match; **UIS,** unique ion signature

Colour Online: See the article online to view Figs 2, 4 and 5 in colour.

While it is basic, accurate determination of monoisotopic masses is non-trivial. The main cause is the interference of other peaks, which is common in mass spectra, because noise peaks and co-eluted precursors are common in the precursor isolation window [6]. In this case, incorrect monoisotopic mass-to-charge ratios (m/z) or charge states may be exported, which occurs in some software, e.g. BioWorks. Furthermore, co-eluted precursors may produce fragment ions in the corresponding MS/MS spectra (mixed spectra) and could be identified in a database search, if the monoisotopic peak of each co-eluted precursor should also be exported. In short, the interference from other peaks, including co-eluted precursors, should be carefully treated in a determination method of monoisotopic peaks.

For the purpose of determining monoisotopic peaks, about four methods have been reported. Among them is the well-known averagine model [7], in which an averaged molecular formula is obtained from a protein database and used to estimate the peptide molecular formula, and then the distance between the experimental and the estimated isotopic distribution is calculated to determine the monoisotopic mass. This idea is used in many software tools, e.g. THRASH [8], Decon2LS [9], DeconMSn [10], Hardklor [11], Bullseye [12], and DTASuperCharge [13]. Similar to the averagine model, another method is based on the relationship between the intensity ratio of adjacent isotopic peaks and the peptide mass [14–16], in which the theoretical relationship is obtained from a protein database, and then the distance between the experimental and the theoretical relationship is calculated to determine the monoisotopic mass. Both methods determine the monoisotopic mass based on a single MS scan, in which low-abundance precursors are sometimes selected at the two ends of elution profiles. This becomes problematic when these precursors exhibit unusual isotopic distributions that are dissimilar to the theoretical ones.

To overcome the disadvantage of low-abundance precursors, many software tools consider the elution profile of a peak over several MS scans, e.g. MaxQuant [17], Raw2MSM [18], VIPER [19], Superhirn [20], MapQuand [21], msInspect [22], Peplist [23], and MZmine [24], most of which are described in the review [25]. However, the three types of methods just mentioned (averagine model, intensity ratio, and elution profile) seldom consider the interference of other peaks. They export only one candidate monoisotopic peak for each MS/MS spectrum. Therefore, they may miss some correct monoisotopic peaks because of the interference of other peaks. The fourth method strives to include the correct monoisotopic peaks by searching the MS/MS spectra with a large precursor mass tolerance, e.g. ± 3.1 Da, and filtering MS/MS spectra with a small precursor mass tolerance, e.g. ± 10 ppm, in each local region that corresponds to precursor mass errors of 0, 1, 2, and 3. When the charge state of the precursor is incorrect, this strategy will also miss the correct monoisotopic peaks. Furthermore, co-eluted precursors are out of consideration in all of these four methods.

Recently, several methods have tried to consider co-eluted precursors. Some export the monoisotopic peaks of co-eluted precursors separately in the isolation window for identification, while the MS/MS spectrum is the same as the original one [26–29]. Some exclude the fragment ions of the first identified peptide from the MS/MS spectrum and use the left fragment ions to identify the second peptide [30]. Some use the identification idea of cross-linked peptides to identify mixed spectra [31]. Others use simulated mixed spectra to study the influence of co-eluted precursors on database and spectral library searches [32–34]. These methods have shown that mixed spectra are common but less likely to result in accurate identification. Therefore, how to identify mixed spectra effectively is still a problem.

In this paper, we propose a new method, named pParse, to accurately determine the monoisotopic masses of precursors. Because of the existence of co-eluted precursors, pParse exports several probable monoisotopic peaks for precursors. The key point is how to use the relationship between the position of the highest peak and the mass of the first peak to detect candidate clusters. To avoid enumerating all possibilities, we extract three features for each cluster: (i) the sum of the intensity, (ii) the similarity of the experimental and the theoretical isotopic distribution, and (iii) the similarity of elution profiles. pParse uses these features to sort all possible monoisotopic peaks in the precursor isolation window including the co-eluted precursors. The top-ranked monoisotopic peaks are exported and assigned to the corresponding MS/MS spectrum separately. To improve the identification rate of mixed spectra, a new scoring function is designed, which is similar to the identification of selected-reaction monitoring (SRM) data [35]. Concisely, we have achieved two improvements for determining monoisotopic peaks: the recall of correct monoisotopic peaks at high precision and the identification rate of co-eluted precursors.

In the following section, we will introduce the workflow of pParse and the scoring function for the identification of co-eluted precursors. Then we will list some results of improvements and discuss: (i) how to evaluate the correctness of monoisotopic peaks, (ii) why there are incorrect monoisotopic peaks exported, (iii) how to effectively identify co-eluted precursors and what the influence of co-eluted precursors is, (iv) what the difference between pParse and BioWorks is, and (v) what the time and space cost of pParse is.

2 Materials and methods

2.1 Methods

Because the instrument software may determine the monoisotopic peak simply by the highest intensity, a more

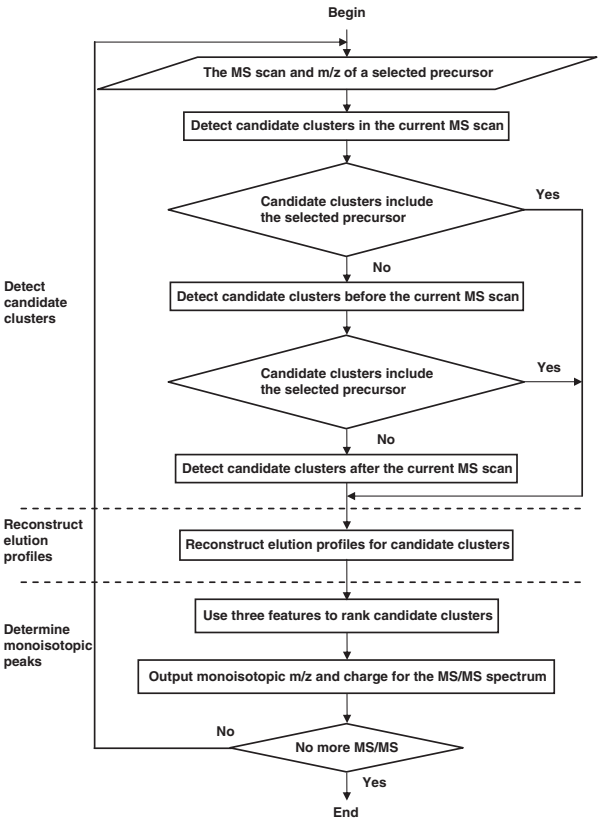

Figure 1. Flow chart of pParse. pParse has three main steps: (a) detecting candidate isotopic clusters in a single MS scan, (b) reconstructing elution profiles for each candidate cluster, and (c) determining monoisotopic peaks by the rank of all clusters. To detect candidate clusters, two factors are considered in each cluster: (i) the relationship between the position of the highest peak and the mass of the first peak and (ii) the similarity of the experimental and the theoretical isotopic distribution. To rank all clusters, three features can be extracted for each cluster: (i) the sum of the intensity, (ii) the similarity of the experimental and the theoretical isotopic distribution, and (iii) the similarity of elution profiles.

accurate method for determining monoisotopic peaks is required after the raw data are produced. Our software tool, pParse, acts as a post-data acquisition procedure to accurately determine the monoisotopic peaks for precursors. The flow chart for pParse, shown in Fig. 1, consists of three major steps: (i) detecting candidate isotopic clusters in a single MS scan, (ii) reconstructing elution profiles for each candidate cluster, and (iii) determining monoisotopic peaks. Each step is described below.

2.1.1 Detecting candidate isotopic clusters in a single MS scan

In the precursor isolation window of the MS scan just preceding an MS/MS scan, candidate isotopic clusters can be detected by scanning peaks from low m/z to high m/z with the assumed m/z interval. When the peak in the isolation window has a low signal-to-noise ratio (S/N), candidate clusters may not include the precursor originally selected for the MS/MS scan. In this case, candidate clusters need to be detected in MS scans before and after the current MS scan (as shown in Fig. 1).

During the candidate detection step, the S/N is computed for each peak to discard noise peaks. The distribution of the peak intensity in the MS scan can be obtained. Then the intensity with the highest frequency is defined as the noise level, and the S/N for a peak is defined as the ratio of the peak intensity to the noise level. MS peaks whose S/Ns are <1 (below the noise level) will be discarded.

Two important criteria are then applied to detect candidate clusters in the precursor isolation window. One is that adjacent peaks in a candidate cluster should have a suitable m/z difference, e.g. 1.0032 Da (the average value of the mass difference of adjacent isotopic peaks obtained from pre-identified MS/MS spectra) divided by the assumed charge state of the precursor ion, which ranges from 2 to 7. The other is that the similarity of the experimental and the theoretical (by the averagine model) isotopic distribution should satisfy given conditions. For example, there are two conditions that must be satisfied if only one candidate cluster is exported starting from the first peak to the end peak. Those two conditions are: (i) the first peak in the current cluster is the highest, and its mass is <1800 Da and (ii) the similarity of the experimental and the theoretical isotopic distribution is more than 0.99 (the threshold can be obtained from pre-identified MS/MS spectra). Otherwise, when the first condition is satisfied, but the second is not satisfied, two candidate clusters are exported: one starts from the first peak to the end peak; the other starts from the second peak to the end peak.

In the second criterion, pParse considers the relationship between the position of the highest peak and the mass of the first peak, which is missed in other averagine model-based methods. The relationship can be inferred from the averagine model: when the peptide mass is <1800 Da, the first peak is the highest; when the peptide mass is between 1800 and 3300 Da, the second peak is the highest. As a result, a list of candidate clusters in the original precursor isolation window can be obtained, including co-eluted precursors.

2.1.2 Reconstructing elution profiles for each candidate cluster

For all peaks in each isotopic cluster, the elution profiles are similar because all of the isotopic peaks are concomitant. But a noise peak does not have an elution profile. Therefore, elution profiles can be used to remove noise peaks. To reconstruct the elution profiles, pParse starts from the MS scan n from which an MS/MS scan is triggered, identifies the peaks belonging to a candidate cluster, and then searches for matching peaks with mass deviations no more than a pre-defined threshold in MS scans $n-1$ and $n+1$. The

matching procedure continues to the neighboring MS scans in both directions until no matching peak is found in two (or other user-defined values) consecutive MS scans. Noise peaks that cannot be assembled into elution profiles are filtered out. For two adjacent peaks in a candidate cluster, the correlation between their elution profiles is calculated using the cosine of the angle between the vectors corresponding to the elution profiles. The correlation of the first two peaks is defined as the similarity of elution profiles in each isotopic cluster.

2.1.3 Determining monoisotopic peaks

After the elution profile's reconstruction, three features can be extracted: (i) the sum of the intensity for each cluster on the current MS scan, (ii) the similarity of the experimental and the theoretical isotopic distribution for each cluster, and (iii) the similarity of elution profiles in each cluster. The ranks of each feature are multiplied as the final scoring function. The top k, e.g. 5, monoisotopic peaks are exported.

After the above three determinations are verified, the MS/MS spectra are exported with all of the determined precursor m/z values and charge states, while the fragment ions are the same as the original ones. On the basis of the algorithm, pParse is implemented using MATLAB and Python. The user manual and the MATLAB source code of pParse are provided in Supporting Information.

2.2 Mixed spectra identification

Mixed spectra are less likely to identify with common database search engines [36–39] because some fragment ions of co-eluted precursors may be low and not fragmented well. To improve the identification rate of mixed spectra, the SRM identification method can be used which uses only one precursor and two fragment ions to identify a peptide [35]. Here, we use ppm mass tolerance to obtain dozens of candidate peptides. For each peptide we follow the steps: (i) calculating the theoretical fragment ions, (ii) counting the fragment ion pairs which occur only in the current peptide (unique ion signature, UIS) and the frequency of non-UIS fragment ions which occur in all candidate peptides, (iii) matching the theoretical fragment ions to the MS/MS spectrum, and (iv) summing the matched intensity of UIS fragment ions and the matched intensity of non-UIS fragment ions divided by their frequency as the score (UIS score). Finally, the peptide of the highest score is exported as the PSM result. To estimate the false discovery rate (FDR), we also calculate the UIS score for the reversed sequence of each candidate peptide [40].

2.3 Data sets

To demonstrate the benefit of pParse, we showed the analysis of two published data sets with high precursor mass accuracy. The first data set was generated from yeast samples, referred to as Yeast data hereafter [41]. The second data set was generated from HeLa cells, referred to as HeLa data hereafter [42]. In the Yeast data, peptides were separated with the LC-MS analysis. In the HeLa data, the digested human cell samples were fractionated with isoelectric focusing (IEF), followed by LC-MS analysis of each fraction. The difference of these two data sets is the separation method. If the samples are only separated by LC, more mixed spectra will occur. If the samples are fractionated with IEF or SDS-PAGE, less mixed spectra will occur but still reach about 50%. The detailed information for these two data sets is shown in Supporting Information Table 1.

3 Results and discussion

3.1 Evaluation for the correctness of monoisotopic peaks

To evaluate the correctness of monoisotopic peaks, we should first generate a confident test set of correct monoisotopic peaks. The central peak in the precursor isolation window can be exported, which is recorded in the raw data. Because the central peak may be the isotopic peak rather than the monoisotopic peak, a large precursor mass tolerance, e.g. ± 3.1 Da, is used to search the exported MS/MS spectra (the searching parameters are shown in Supporting Information Table 2). In each local region that corresponds to precursor mass errors of 0, 1, 2, and 3, search results are filtered with a small precursor mass tolerance, e.g. ± 10 ppm, using the target-decoy strategy.

Second, we can use several kinds of software to export the monoisotopic peaks, such as pParse, MaxQuant, and the instrument software BioWorks (the exporting parameters are shown in Supporting Information Table 3). For each identified monoisotopic peak, we compare its m/z value and charge state with the exported precursors. When the charge states are the same and the difference of m/z values is no more than 10 ppm, the exported precursor is correct, which means the software exports the correct monoisotopic peak. Otherwise, the exported precursor is incorrect.

Third, we sort the identified monoisotopic peaks by their database search scores in a descending order. The higher the database search score, the more reliable the monoisotopic peak. Therefore, we can calculate the two evaluation measurements, precision and recall. Precision is calculated by the number of correct precursors in the current set divided by the size of the current set, e.g. the first k sorted monoisotopic peaks. Recall is calculated by the number of correct precursors in the current set divided by the number of all identified monoisotopic peaks, which is the same measurement as sensitivity.

The above three steps are used to evaluate the correctness of the monoisotopic peak for the cluster of the central peak in the precursor isolation window. To evaluate the correct-

ness of all monoisotopic peaks in the precursor isolation window, we should modify the first step above. In the precursor isolation window, adjacent peaks with a suitable m/z difference constitute a candidate cluster. All the peaks in each candidate cluster are exported (by so called brute force). A small precursor mass tolerance, e.g. ± 10 ppm, is used to search the exported MS/MS spectra. Search results are filtered by the target-decoy strategy. The second and third steps remain the same as the steps above.

The evaluation results for the Yeast data and the HeLa data are shown in Fig. 2. In Fig. 2, the search engine is pFind. Actually, the evaluation approaches can be used with any search engine. Supporting Information Fig. 1 repeats similar results by MASCOT. Concisely, pParse is the most sensitive of all cases, and the sensitivity on all identified monoisotopic peaks reaches more than 98%. In the case of evaluating the correctness of all monoisotopic peaks in the precursor isolation window, the precision of pParse, MaxQuant, and BioWorks was 99–99.1%, 88–93.2%, and 79.9–94.7% at the same recall, and the recall of pParse, MaxQuant, and BioWorks was 98–98.8%, 0.5–17%, and 1.8–36.5% at the same precision (as shown in Table 1), because MaxQuant and BioWorks do not export co-eluted precursors.

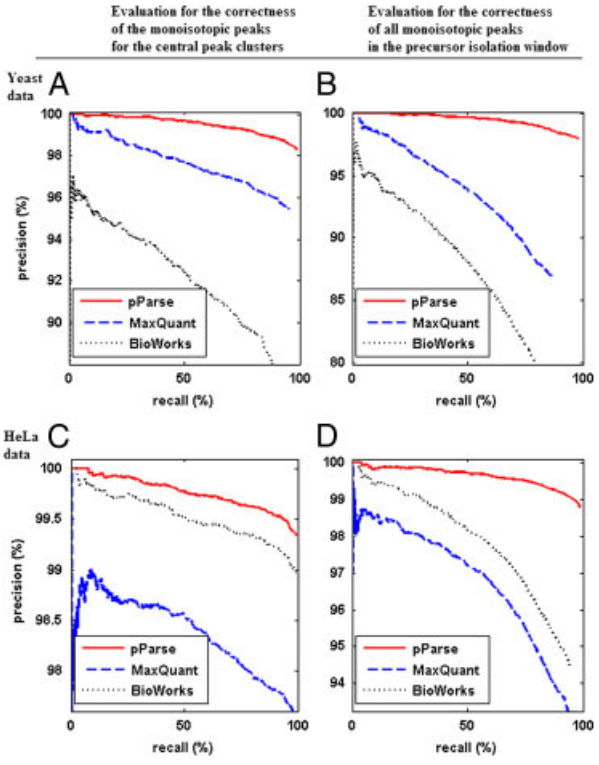

Figure 2. Evaluation for the correctness of monoisotopic peaks. Two evaluation approaches are designed in the text: (i) evaluation for the correctness of the monoisotopic peaks for the central peak clusters and (ii) evaluation for the correctness of all monoisotopic peaks in the precursor isolation window. Points (A) and (B) are the precision recall curves for the Yeast data, while (C) and (D) are the precision recall curves for the HeLa data. In four cases pParse is the most sensitive and the sensitivity on all identified monoisotopic peaks reaches more than 98%.

In the two cases of evaluating the correctness of monoisotopic peaks, there are still a few precursors exported by pParse that are different from the identified monoisotopic peaks. In fact, the identified monoisotopic peaks are indeed incorrect, because the corresponding monoisotopic peaks to the peptides are missing in the precursor isolation window. After the deamidation modification is considered, the peptides can be identified with this modification and the corresponding monoisotopic peaks appear in the precursor isolation window, which are the same as the monoisotopic peaks exported by pParse. Checking the identification in this way, it is found that 99% of the identified monoisotopic peaks are correct. Therefore, this approach has proven suitable for evaluating the correctness of monoisotopic peaks.

3.2 The reasons for exporting incorrect monoisotopic peaks

After evaluation with the identified monoisotopic peaks, we can find out some reasons for BioWorks and MaxQuant exporting incorrect monoisotopic peaks. BioWorks has shown a propensity for exporting incorrect monoisotopic peaks in two cases (Supporting Information Fig. 2). The first case is when the central peak is the highest in the cluster (but not the first peak), the central peak is exported as the monoisotopic peak; the second case is that when another cluster appears before the central peak cluster with the same m/z difference, one peak in the former cluster is exported as the monoisotopic peak. The first case is the main cause, because BioWorks seems to determine monoisotopic peaks by the highest intensity rather than the isotopic distribution.

In five cases MaxQuant is easy to export incorrect monoisotopic peaks: (i) when co-eluted precursors occur, only one monoisotopic peak is exported; (ii) when there is another cluster before the central peak cluster with the same m/z difference, one peak in the former cluster is exported as the monoisotopic peak; (iii) when the central peak cluster has two candidate charge states and the clusters of both charge states have elution profiles, only one charge state is exported; (iv) when there are two peaks close to the monoisotopic peak, the mass deviation of the exported monoisotopic peak is a little larger; (v) a few precursors are filtered out by MaxQuant. The first four cases are shown in Supporting Information Fig. 3. The first three cases are the main causes. The first two cases occur because there are co-eluted precursors that need be considered in monoisotopic peak detection. The third case occurs because elution profile cannot distinguish them in the isolation window, but the isotopic distribution can help.

In the fourth case of MaxQuant, pParse may also export the monoisotopic peak with a little larger mass deviation, because the monoisotopic peak has been interfered with a

Table 1. Comparison of three software programs' precision at the same recall and recall at the same precision to evaluate the correctness of all monoisotopic peaks in the precursor isolation window

	Precision of BioWorks (%)	Precision of MaxQuant (%)	Precision of pParse (%)
Recall of 79.9% in the Yeast data	79.9	88	99
Recall of 93.2% in the HeLa data	94.7	93.2	99.1
	Recall of BioWorks (%)	Recall of MaxQuant (%)	Recall of pParse (%)
Precision of 98% in the Yeast data	1.8	17	98
Precision of 98.8% in the HeLa data	36.5	0.5	98.8

nearby peak. But pParse can deal well with other four cases of MaxQuant and the two cases of BioWorks. Therefore, pParse outperforms MaxQuant and BioWorks.

3.3 The identification and the influence of co-eluted precursors

Using pParse we find that in all MS/MS spectra, the proportion of mixed spectra (at least two precursors in the isolation window) is 50–60%. Unfortunately, mixed spectra are less likely to result in identification with common database search engines, because some fragment ions for the co-eluted precursors may be low and not fragmented well. For example, using the search engine pFind we identified 5–8% of all co-eluted precursors at the FDR$\leq$1%. Then we design a new scoring function (UIS score) similar to the SRM identification. To estimate the FDR, we also calculate the UIS score for the reversed sequence of each candidate peptide. We searched the Yeast data on *E. coli* database with UIS and found that the ratio of the number of co-eluted precursors identified from target and decoy peptides was 0.987:1, which means the estimated FDR is higher than the real FDR. Therefore, we set the FDR$\leq$5% for UIS. Furthermore, because with pFind the ratio was 1.09:1, we set the FDR$\leq$1% for pFind.

In the Yeast data, UIS identified 1343 co-eluted precursors as having the same peptides as pFind, and four co-eluted precursors as having different peptides from pFind. pFind identified 415 more co-eluted precursors which were missed by UIS, and UIS identified 1781 more co-eluted precursors which were missed by pFind (as shown in Fig. 3). We checked the 1781 more co-eluted precursors and found that 70% of them were the same top-one peptides found by pFind which were filtered out by the FDR$\leq$1%. After merging the results of UIS and pFind, 16% of all co-eluted precursors were identified. Therefore, UIS can improve the identification rate of co-eluted precursors.

A typical example of mixed spectrum is shown in Fig. 4. Peptide A of the central peak cluster is identified both by pFind and UIS. The co-eluted peptide B is only identified by UIS. From the matched MS/MS spectrum, we know that most of the fragment ions of peptide A are high and consecutive, while only a few fragment ions of peptide B are high or consecutive. Because pFind gives high weights to consecutive ions, the score of peptide A is much higher than that of peptide B. Thus, peptide A is easy to be identified by

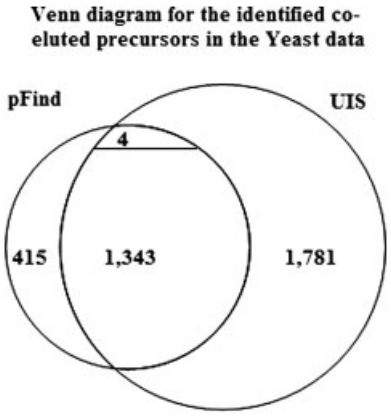

Figure 3. Venn diagram for the identified co-eluted precursors in the Yeast data. The left smaller circle represents the number of identified co-eluted precursors by pFind. The right larger circle represents the number of identified co-eluted precursors by UIS. UIS identified 1343 co-eluted precursors with the same peptides as pFind, and four co-eluted precursors with different peptides from those found by pFind. pFind identified 415 more co-eluted precursors, and UIS identified 1781 more co-eluted precursors.

pFind, but peptide B may be unidentified. Anyhow, there are still a few high peaks that cannot be matched by peptide A. When these few high peaks can only be matched with peptide B, we can also identify peptide B. UIS uses the uniqueness of two peak pairs to identify less consecutive fragment ions. Therefore, the uniqueness of UIS and the consecutiveness of pFind are complementary and can be combined to improve the identification rate of mixed spectra.

After the combinational identification of mixed spectra by UIS and pFind, more peptides and proteins can be identified. The number of identified mixed spectra in the Yeast data is shown in Table 2. In the Yeast data, 26.2% of all identified peptides were exclusively from co-eluted peptides, and 35.8% of all identified proteins were exclusively from co-eluted peptides (one example shown in Fig. 5A). About 30% of all identified proteins had been identified from the central peak precursors and the coverage values were increased by the co-eluted precursors (one example shown in Fig. 5B). The detailed protein coverage of the Yeast data is shown in Supporting Information, which shows that the average protein coverage identified from the central peak precursors was 18.8%, and the average protein coverage identified from

the co-eluted precursors was 3.2%. Therefore, the detection and identification of co-eluted precursors can increase the number of identified proteins and protein coverage greatly.

3.4 Comparison of pParse with the instrument software

Though the instrument software, BioWorks, can export some correct monoisotopic peaks, there are still some cases in which BioWorks exports incorrect monoisotopic peaks. One way to resolve the problem is by searching MS/MS spectra exported by BioWorks with a large precursor mass tolerance and filtering MS/MS spectra with a small precursor mass tolerance in each local region, e.g. 0, 1, 2, and 3. In the Yeast data, this strategy identified 5356 peptides. Then we search the MS/MS spectra exported by pParse with a small precursor mass tolerance. The second strategy identified 7668 peptides. The overlapping of these two strategies is 5339. The second strategy identified 2329

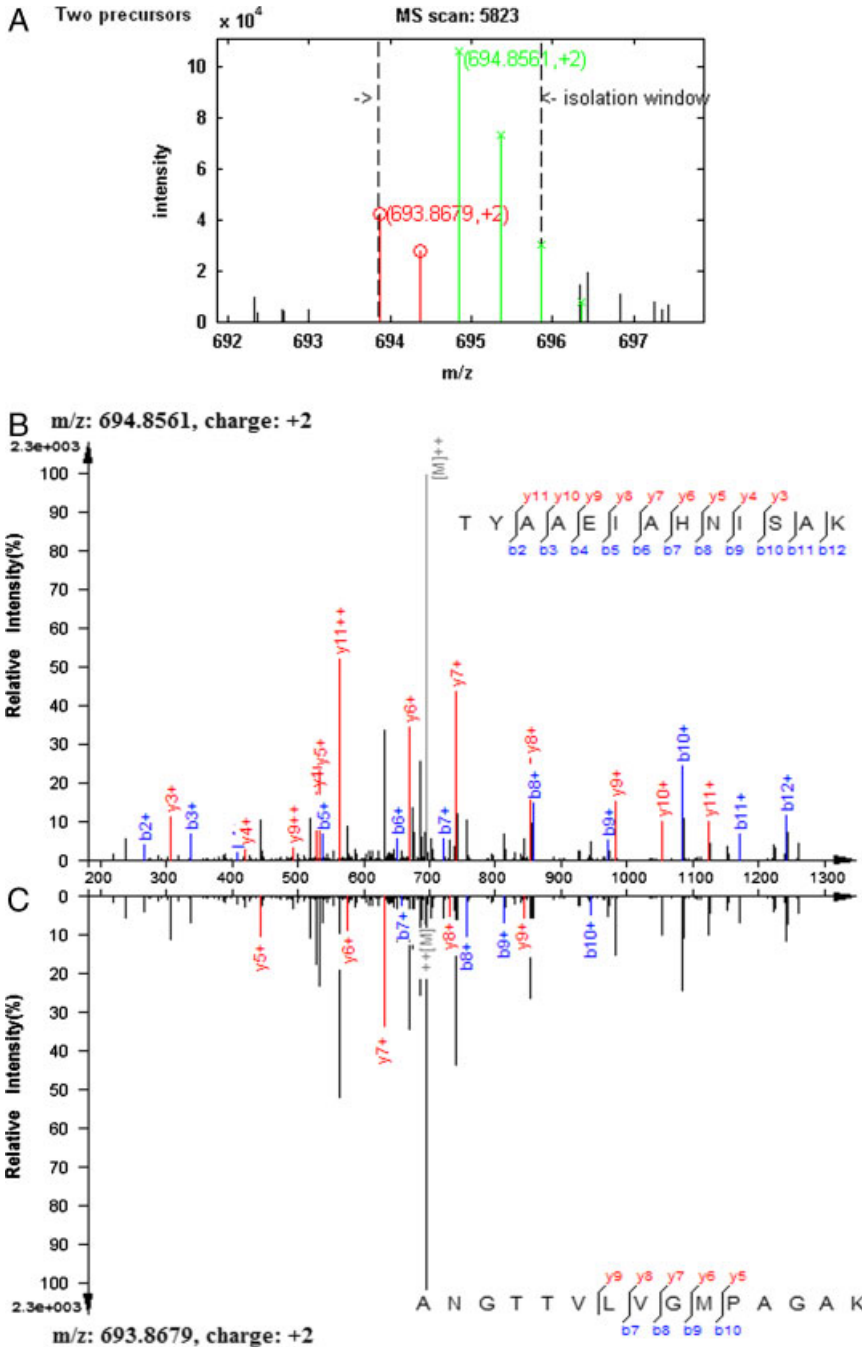

Figure 4. An example of an identified mixed spectrum. (A) Two monoisotopic peaks are determined in the precursor isolation window: one is the central peak; the other is the peak close to the left side of the isolation window. (B) The peptide A of the central peak precursor is matched with the MS/MS spectrum. (C) The peptide B of the co-eluted precursor is matched with the MS/MS spectrum. Most of the fragment ions of peptide A are high and consecutive, while a few fragment ions of peptide B are high or consecutive. Actually, peptide B is the top-one PSM in pFind, but it is below the FDR threshold. UIS gives peptide B a relatively high score and identifies it. UIS and pFind can be combined to give more reliable peptides.

more peptides, and the first strategy identified only 17 more peptides (as shown in Fig. 6). Two cases occur for these 17 peptides in pParse: (i) the mass deviation of the exported monoisotopic peak is a little larger and (ii) there is a deamidation modification. The 2329 peptides only identified by pParse are mainly from co-eluted precursors. Therefore, pParse outperforms BioWorks in determining monoisotopic peaks.

3.5 The performance of pParse

In the precursor isolation window, adjacent peaks with a suitable m/z difference constitute a candidate cluster. All the peaks in each candidate cluster can be exported, which is the so-called brute force. In this way, correct monoisotopic peaks cannot be missed. We compared the exporting time and the number of exported precursors for brute force, pParse, MaxQuant, and BioWorks in the Yeast data. The result is shown in Table 3. (All the programs were executed on the same PC: Intel Core 2 Duo processor 2.66 GHz, 2 GB RAM, Windows XP OS.) The exporting time of all four ways is acceptable, while the difference is the sensitivity of correct monoisotopic peaks and the number of exported monoisotopic peaks. Because pParse considers the most probable monoisotopic peaks and avoids brute force, the sensitivity of pParse is similar to that of brute force. The number of exported precursors of pParse is two times more than that of BioWorks, while the number of exported precursors of brute force is six times more than that of BioWorks. Therefore, pParse reaches high sensitivity and controls the number of exported precursors, whereas sensitivity is much more important in monoisotopic peak determination.

Table 2. The number of mixed spectra according to the number of identified peptides in the Yeast data by UIS and pFind

Number of identified peptides in each MS/MS spectra	2	3	4
Number of mixed spectra	1765	91	3

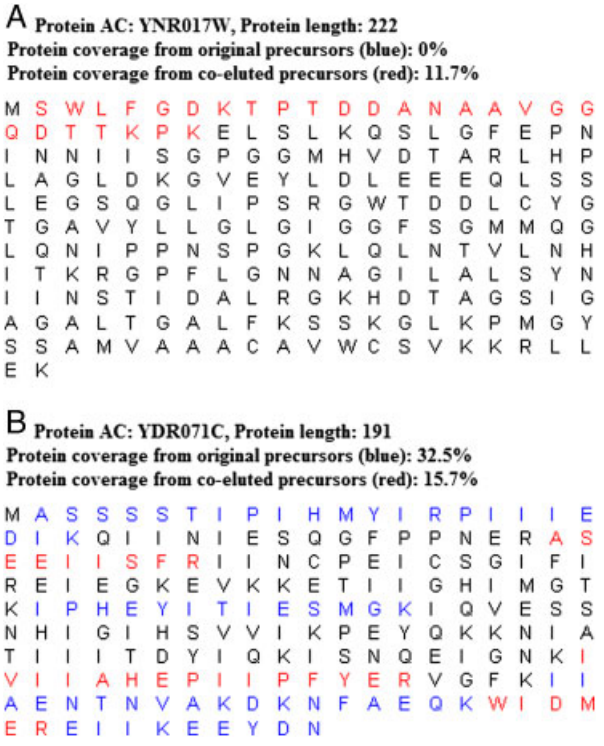

Figure 5. The influence of co-eluted precursors. Co-eluted precursors can increase the protein coverage. In the Yeast data, 735 proteins were solely identified from co-eluted precursors (e.g. A); 631 proteins had been identified from central peak precursors and the coverage values were increased by co-eluted precursors (e.g. B); 685 proteins were solely identified from central peak precursors.

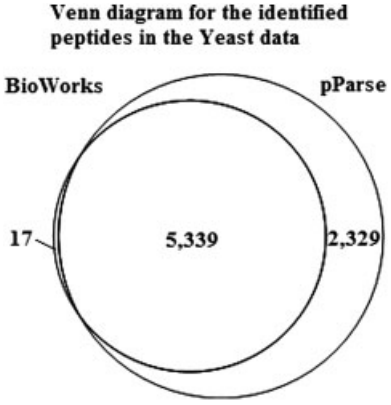

Figure 6. Venn diagram for the identified peptides in the Yeast data. The left smaller circle represents the number of identified peptides from BioWorks. The right larger circle represents the number of identified peptides from pParse. The first strategy is that MS/MS spectra are exported by BioWorks, searched with a large precursor mass tolerance, and filtered with a small precursor mass tolerance in each local region that corresponds to precursor mass errors of 0, 1, 2, and 3. The second strategy is that MS/MS spectra are exported by pParse and searched with a small precursor mass tolerance.

Table 3. Comparison of the exporting time and the number of exported precursors for brute force, pParse, MaxQuant, and BioWorks in the Yeast data

Tools	The exporting time (min)	The number of exported precursors
Brute force	33.8	237 452
pParse	33.7	88 419
MaxQuant	44.6	45 760
BioWorks	21.4	39 829

4 Concluding remarks

Monoisotopic peak determination and co-eluted precursor identification are challenges in interpreting mass spectra. In this paper, we presented pParse, a new method to determine the monoisotopic masses of precursors for MS/MS spectra. Because pParse uses a new method to detect candidate clusters and three important features to sort them, the sensitivity of pParse reaches more than 98%. Though co-eluted precursors are less likely to identify, we use the uniqueness of UIS and the consecutiveness of pFind to improve the identification rate of mixed spectra from 8 to 16%, and increase protein identification and coverage greatly.

pParse is designed for Thermo FT/Orbitrap RAW files, i.e. high-resolution RAW files for shotgun proteomics and is not suitable for low-resolution RAW files. It is known that other mass spectrometers also need to determine correct monoisotopic peaks and detect co-eluted precursors. Therefore, we will extend the algorithm of pParse to these mass spectrometers as a future work.

The authors thank Meng-Qiu Dong, Xiao-Hong Qian, Wei Jia, Zhuang Lu, Shen-Heng Guan, Tao Xu, Ding Ye, Yan-Jie Wu, Sheng-Bo Fan, Kun He, and Long Wu for valuable discussions, reviewers for valuable suggestions, Dr. Tabb and Dr. Cox for providing the RAW format data. This work was supported by the National Key Basic Research & Development Program (973) of China under Grant Nos. 2010CB912701 and 2012CB910602; by the CAS Knowledge Innovation Program under Grant No. KGGX1-YW-13; by the National Natural Science Foundation of China under Grant No. 30900262; and by the National High Technology Research and Development Program (863) of China under Grant Nos. 2007AA02Z315 and 2008AA02Z309.

The authors have declared no conflict of interest.

5 References

[1] Aebersold, R., Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, *422*, 198–207.

[2] Nesvizhskii, A. I., Vitek, O., Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 2007, *4*, 787–797.

[3] Marshall, A. G., Hendrickson, C. L., Jackson, G. S., Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* 1998, *17*, 1–35.

[4] Hu, Q., Noll, R. J., Li, H., Makarov, A. et al., The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* 2005, *40*, 430–443.

[5] Strittmatter, E. F., Rodriguez, N., Smith, R. D., High mass measurement accuracy determination for proteomics using multivariate regression fitting: application to electrospray ionization time-of-flight mass spectrometry. *Anal. Chem.* 2003, *75*, 460–468.

[6] Michalski, A., Cox, J., Mann, M., 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.* 2011, *10*, 1785–1793.

[7] Senko, M. W., Beu, S. C., Mclafferty, F. W., Determination of Monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* 1995, *6*, 229–233.

[8] Horn, D. M., Zubarev, R. A., McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* 2000, *11*, 320–332.

[9] Jaitly, N., Mayampurath, A., Littlefield, K., Adkins, J. N. et al., Decon2LS: an open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics* 2009, *10*, 87.

[10] Mayampurath, A. M., Jaitly, N., Purvine, S. O., Monroe, M. E. et al., DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics* 2008, *24*, 1021–1023.

[11] Hoopmann, M. R., Finney, G. L., MacCoss, M. J., High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.* 2007, *79*, 5620–5632.

[12] Hsieh, E. J., Hoopmann, M. R., MacLean, B., MacCoss, M. J., Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* 2010, *9*, 1138–1143.

[13] Mortensen, P., Gouw, J. W., Olsen, J. V., Ong, S. E. et al., MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J. Proteome Res.* 2010, *9*, 393–403.

[14] Zhang, X., Hines, W., Adamec, J., Asara, J. M. et al., An automated method for the analysis of stable isotope labeling data in proteomics. *J. Am. Soc. Mass Spectrom.* 2005, *16*, 1181–1191.

[15] Park, K., Yoon, J. Y., Lee, S., Paek, E. et al., Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. *Anal. Chem.* 2008, *80*, 7294–7303.

[16] Zhang, J., Gao, W., Cai, J., He, S. et al., Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2005, *2*, 217–230.

[17] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 2008, *26*, 1367–1372.

[18] Olsen, J. V., de Godoy, L. M., Li, G., Macek, B. et al., Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* 2005, *4*, 2010–2021.

[19] Monroe, M. E., Tolic, N., Jaitly, N., Shaw, J. L. et al., VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* 2007, *23*, 2021–2023.

[20] Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S. et al., SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 2007, *7*, 3470–3480.

[21] Leptos, K. C., Sarracino, D. A., Jaffe, J. D., Krastins, B., Church, G. M., MapQuant: open-source software for large-scale protein quantification. *Proteomics* 2006, *6*, 1770–1782.

[22] Bellew, M., Coram, M., Fitzgibbon, M., Igra, M. et al., A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* 2006, *22*, 1902–1909.

[23] Li, X. J., Yi, E. C., Kemp, C. J., Zhang, H., Aebersold, R., A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics* 2005, *4*, 1328–1340.

[24] Katajamaa, M., Miettinen, J., Oresic, M., MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 2006, *22*, 634–636.

[25] Zhang, J., Gonzalez, E., Hestilow, T., Haskins, W., Huang, Y., Review of peak detection algorithms in liquid-chromatography–mass spectrometry. *Curr. Genomics* 2009, *10*, 388–401.

[26] Scherl, A., Tsai, Y. S., Shaffer, S. A., Goodlett, D. R., Increasing information from shotgun proteomic data by accounting for misassigned precursor ion masses. *Proteomics* 2008, *8*, 2791–2797.

[27] Luethy, R., Kessner, D. E., Katz, J. E., Maclean, B. et al., Precursor-ion mass re-estimation improves peptide identification on hybrid instruments. *J. Proteome Res.* 2008, *7*, 4031–4039.

[28] Carvalho, P. C., Xu, T., Han, X., Cociorva, D. et al., YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics* 2009, *25*, 2734–2736.

[29] Zhang, N., Li, X. J., Ye, M., Pan, S. et al., ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* 2005, *5*, 4096–4106.

[30] Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A. et al., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* 2011, *10*, 1794–1805.

[31] Wang, J., Bourne, P. E., Bandeira, N., Peptide identification by database search of mixture tandem mass spectra. *Mol. Cell. Proteomics* 2011, *10*, M111.010017.

[32] Alves, G., Ogurtsov, A. Y., Kwok, S., Wu, W. W. et al., Detection of co-eluted peptides using database search methods. *Biol. Direct* 2008, *3*, 27.

[33] Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K. et al., Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* 2010, *9*, 4152–4160.

[34] Wang, J., Perez-Santiago, J., Katz, J. E., Mallick, P., Bandeira, N., Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics* 2010, *9*, 1476–1485.

[35] Sherman, J., McKay, M. J., Ashman, K., Molloy, M. P., Unique ion signature mass spectrometry, a deterministic method to assign peptide identity. *Mol. Cell. Proteomics* 2009, *8*, 2051–2062.

[36] Eng, J. K., Mccormack, A. L., Yates, J. R., An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[37] Fu, Y., Yang, Q., Sun, R., Li, D. et al., Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* 2004, *20*, 1948–1954.

[38] Wang, L. H., Li, D. Q., Fu, Y., Wang, H. P. et al., pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 2007, *21*, 2985–2991.

[39] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[40] Elias, J. E., Haas, W., Faherty, B. K., Gygi, S. P., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* 2005, *2*, 667–675.

[41] Tabb, D. L., Ma, Z.-Q., Martin, D. B., Ham, A.-J. L., Chambers, M. C., DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.* 2008, *7*, 3838–3846.

[42] Cox, J., Mann, M., Is proteomics the new genomics? *Cell* 2007, *130*, 395–398.

