True ion pick (TIPick): a denoising and peak picking algorithm to extract ion signals from liquid chromatography/mass spectrometry data

Tsung-Jung Ho,^{a,b†} Ching-Hua Kuo,^{a,c†} San-Yuan Wang,^{a,b}
Guan-Yuan Chen^{a,b} and Yufeng J. Tseng^{a,b,c,d*}

Liquid Chromatography - Time of Flight Mass Spectrometry has become an important technique for toxicological screening and metabolomics. We describe *TIPick* a novel algorithm that accurately and sensitively detects target compounds in biological samples. *TIPick* comprises two main steps: background subtraction and peak picking. By subtracting a blank chromatogram, *TIPick* eliminates chemical signals of blank injections and reduces false positive results. *TIPick* detects peaks by calculating the $S(CC_{INI})$ values of extracted ion chromatograms (EICs) without considering peak shapes, and it is able to detect tailing and fronting peaks. *TIPick* also uses duplicate injections to enhance the signals of the peaks and thus improve the peak detection power. Commonly seen split peaks caused by either saturation of the mass spectrometer detector or a mathematical background subtraction algorithm can be resolved by adjusting the mass error tolerance of the EICs and by comparing the EICs before and after background subtraction. The performance of *TIPick* was tested in a data set containing 297 standard mixtures; the recall, precision and F-score were 0.99, 0.97 and 0.98, respectively. *TIPick* was successfully used to construct and analyze the NTU MetaCore metabolomics chemical standards library, and it was applied for toxicological screening and metabolomics studies. Copyright © 2013 John Wiley & Sons, Ltd.

Supplementary electronic material for this paper is available at the Wiley InterScience website.

Keywords: algorithm; metabolomics; chemometrics; toxicological screening; Liquid Chromatography - Time of Flight Mass Spectrometry

Introduction

Liquid Chromatography - Time of Flight Mass Spectrometry (LC-TOF-MS) provides high mass accuracy that enables the use of exact monoisotopic masses and isotopic patterns for compound identification.^[1] This facilitates the detection of rare chemicals or metabolites for which standard compounds are difficult to obtain. Therefore, LC-TOF-MS has recently become an important technique for high throughput toxicological, metabolomics, drug metabolism and pharmacokinetics studies.^[2–6] For both toxicological screening and metabolomics, accurate and sensitive peak extraction of target compounds is a key step to obtain accurate results.

Chemical and random noises are inevitably generated by commonly used ionization sources in LC-TOF-MS.^[7–9] Chemical noise is mainly contributed by the mobile phase and endogenous sample matrix,^[10] while the random noise is usually caused by thermal variations or the instability of the ionization interface.^[11] Noise obscures the peaks of target compounds and makes the peak picking procedure challenging. To minimize the noise effect, denoising algorithms have been developed that are based on variable selection or background subtraction. Variable selection methods such as CODA,^[10] RINE,^[11] and CODA_DW^[12] were developed to eliminate low-quality extracted ion chromatograms (EICs) to generate total ion chromatograms (TICs) with low background noise. Background subtraction algorithms subtract undesired components from

the data acquired for a sample to avoid the false indication of the presence of component or compounds if such component or compounds can be found in the blank samples (therefore, reducing the false positive results).^[13] Current background subtraction methods include spectral subtraction (often found in commercial software from instrument vendors) and standalone algorithms, e.g. thorough background subtraction^[14] and BgS-NoRA.^[15] Commercial software, including the Qualitative Analysis function of the Agilent MassHunter Workstation Software, requires users to manually select and extract background spectra around the peaks of interest. However, most of

* Correspondence to: Yufeng J. Tseng, Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei, Taiwan. E-mail: yjtseng@csie.ntu.edu.tw

† These authors contributed equally to this work.

a The Metabolomics Core Laboratory, Center of Genomic Medicine, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei, Taiwan, 106

b Department of Computer Science and Information Engineering, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei, Taiwan, 106

c School of Pharmacy, College of Medicine, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei, Taiwan, 106

d Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No. 1 Sec. 4, Roosevelt Road, Taipei, Taiwan, 106

the time, the chemical noise cannot be effectively removed manually. In contrast, standalone algorithms analyze all ions in the blank injection (within a specified time window around the analyte scan) for potential subtraction of ions found in the samples. These algorithms then remove the background ions and overcome the problem of both retention time (RT) shifts and mass-to-charge ratio (m/z) errors in the LC-MS by setting a tolerance for the RT shift and the m/z error. However, background subtraction algorithms can falsely remove signals and generate artificial split peaks that complicate subsequent data analysis if the parameters of the algorithms are not selected appropriately.

Peak picking algorithms such as *matchedFilter* and *centWave* from XCMS^[16,17] and peak detection algorithms from the MZmine2^[18] and MAVEN^[19] software packages, pick peaks by incorporating smoothing algorithms (e.g. Gaussian filtering) to reduce random noise. These algorithms have been used to effectively detect many small-molecule compounds; however, false negative results can be found if the target peaks are not of Gaussian shapes.

To overcome the limitations of current algorithms, we have developed a novel denoising and peak picking algorithm, *TIPick*, for target analysis to detect true peaks in LC-MS data. The performance of *TIPick* was evaluated with a set of 297 standards from the NTU MetaCore LC-MS library and with a set of human urine-containing forensic drugs and a metabolomics study.

Theory

Overall procedures of the main algorithm

TIPick comprises two major steps: background subtraction and peak picking. Background subtraction removes the chemical noise in both the m/z and the chromatographic time domains. After background subtraction, EICs are generated by *TIPick* based on the target m/z values. *TIPick* automatically calculates the protonated or deprotonated m/z value of the target molecular weight depending on the ionization mode that the user specifies. In the peak picking process, peak enhancement and binning are performed to avoid the influence of random noise. *TIPick* selects peaks in each EIC by recursively detecting all peaks in the EIC based on a scoring function. If the score of an EIC is lower than a specified value (0.75 in our example), *TIPick* proceeds to a peak enhancement step to enhance this particular EIC. After the enhancement step, the score of the EIC is recalculated and additional peaks are selected. The resultant RT is reported by *TIPick* during the peak extraction phase. The overall procedure of *TIPick* is illustrated in Fig. 1.

Background subtraction

In the background subtraction step, centroid accurate LC-MS metabolomics data are converted by an in-house library to the mzList format. mzList contains a data list structure in the R statistical programming language,^[20] where each entry represents a mass scan spectrum. In the following sections, the standard mixture is referred to as 'the analyte file'. The solvent blank for the analysis of the NTU MetaCore metabolomics chemical standards library and the urine blank (no spiking with standards) for the analysis of the forensic drugs are defined as 'the blank files'.

The background subtraction step of *TIPick* is as follows:

- 1) *TIPick* sets an RT window (12 s in our example) for each ion I_b with RT ' t ' and m/z ' x ' in the blank file.
- 2) *TIPick* searches for ion I_a in the same RT window in the analyte file to match ion I_b within a user-defined mass tolerance δ (10 ppm relative mass accuracy in our example), i.e.

$$x(1-10\text{ ppm}) \leq m/z \text{ of } I_a \leq x(1+10\text{ ppm}) \text{ and } t-12\text{ sec} \leq \text{RT of } I_a \leq t+12\text{ s}$$

- 3) *TIPick* removes the detected ion I_a if its intensity value is lower than the intensity value of ion I_b multiplied by a factor corresponding to the specified intensity tolerance ($\text{Int_tolerance} = 1.5$ in our example), i.e. the intensity of $I_a \leq 1.5 \times$ the intensity of I_b .

TIPick background subtraction algorithm was developed independent from both thorough background subtraction^[14] and BgS-NoRA.^[15] The major difference between *TIPick* and BgS background subtraction algorithm is that *TIPick* removes the detected ion I_a while BgS reassigns I_a as I_a minus (I_b multiply a scaling factor). This difference affects the memory size needed while performing the background subtraction step. *TIPick* could utilize much less memory space since all the 'zeros' generated after removal of I_a in the background were omitted from memory storage. Also, *TIPick* background subtraction algorithm allows selection of optimal parameters. In general, background subtraction algorithms can falsely remove signals and generate artificial split peaks that complicate subsequent data analysis if the parameters of the algorithms are not selected appropriately. From our experience, the optimal parameter selection greatly affects the performance of *TIPick*. We further demonstrate the effect of parameters selection in the RESULTS AND DISCUSSION section.

Peak picking

TIPick selects peaks in each EIC by recursively detecting all peaks in the EIC based on a scoring function, $\Psi(\text{CC}_{\text{INI}})$. $\Psi(\text{CC}_{\text{INI}})$ uses CC_{INI} (the Cumulative Curve of the Integral of the Normalized Intensity) to generate a score for each EIC. Peak enhancement is performed as an embedded part of the peak picking process. For convenience, we use $S(\text{CC}_{\text{INI}})$ as the abbreviation of the score of CC_{INI} in the following sections. $S(\text{CC}_{\text{INI}})$ can be used to evaluate the quality of the EIC. The detailed peak picking procedures are as follows:

- Peak enhancement: the peak enhancement step utilizing two criteria – omissions of low-intensity data points and the averaging of duplicate EICs. *TIPick* omits data points with lower intensities than a threshold value – defined as the most intense peak in the EIC multiplied by a user-input intensity factor – and then calculates $S(\text{CC}_{\text{INI}})$. A series of intensity factors (corresponding to thresholds for different peak enhancement conditions) is allowed with the *TIPick* algorithm. In practice, three possible independent enhancement steps could be applied to every EIC to enhance the signal intensities in the EIC. They are direct omission of low-intensity data points before averaging of duplicate EICs, averaging of duplicate EICs, and omission of low-intensity data points after averaging of

Figure 1. (a) Logical flow of the *TIPick* algorithm. (b) Detailed workflow of peak picking with the *TIPick* algorithm. *TIPick* selects peaks in each EIC based on the $S(CC_{INI})$ score.

duplicate EICs (see Fig. 1b). Therefore, user should provide a set of parameters consisting of 'the series of intensity factors before averaging of duplicate EICs' and 'the series of intensity factors after averaging of duplicate EICs'. After peak enhancement, *TIPick* then calculates $S(CC_{INI})$.

- $S(CC_{INI})$ calculation:

a. Intensity normalization: the intensity value of each data point in the EIC is normalized by dividing the total intensity of all ions in the EIC.

b. Peak binning: continuous data points with positive intensity values are grouped into regions. We define continuous data points as those that have positive intensities, and we allow a maximum gap of 2 s between data points. The intensity of the first data point of each region is replaced by the integral intensity of that region, whereas the intensities of the remaining data points are set to zero. The reorganized data points and intensities are then used to generate CC_{INI} .

c. The $S(CC_{INI})$ value is calculated according to the following equations:

$$S(CC_{INI}) = \Psi(CC_{INI}) = \begin{cases} 1 - \frac{\sum_{i=1}^n \sum_{j=1}^i I^*(j)}{0.5 \times \max(n, W_{avg} \times Hz)}, & \text{when } n \geq W_{min} \times Hz \\ 0, & \text{otherwise} \end{cases}$$

In the above equation, $I^*(j)$ is the value of CC_{INI} at the j^{th} point, i is the i^{th} data points after intensity normalization and peak binning, n is the number of data points used to calculate CC_{INI} , Hz is the acquisition rate of the mass spectrometer, W_{avg} is the average peak width (estimated experimentally) and W_{min} is user specified minimum peak width. *TIPick* defines a peak width as the number of continuous data points around the peak apex divided by the acquisition rate of the mass spectrometer. The next step is to begin the recursive peak picking process according to the $S(CC_{INI})$ value.

- Recursive peak picking: The value of $S(CC_{INI})$ can tell us whether there is at least one true peak in an EIC. For each EIC after the peak enhancement, if the $S(CC_{INI})$ value is larger than the threshold, *TIPick* would decide there is at least one true peak (the largest peak) in this EIC and then extract both the RT and intensity of the peak in the next peak extraction step. If the $S(CC_{INI})$ value is zero, the peak picking process is stopped. If the $S(CC_{INI})$ value is less than the SCC_thresh value (0.75 in our example), *TIPick* checks which the previous peak enhancement steps has been performed. If peaks were enhanced by averaging the duplicate EICs followed by the removal of the low-intensity data points (peak enhancement step 3), the recursive peak picking process is then stopped. Otherwise, *TIPick* returns to the peak enhancement procedure. If the previous enhancement was step 1 only (omission of the low intensity data), then the averaging of duplicate EICs (peak enhancement step 2) will be performed. If step 2 was the last performed, *TIPick* then moves on to peak enhancement step 3. After each round of execution, the $S(CC_{INI})$ is recalculated. Once the recursive peak picking process is finished, *TIPick* decides there are no more peaks in the EIC and then moves to the next EIC for peak picking.
- Peak extraction: If the $S(CC_{INI})$ value is greater than the SCC_thresh , at least one peak is considered detected (the largest peak) and added to the peak table of the sample. The RT of the time point with the highest signal intensity of the EIC is selected and reported at this step, and the detected peak is removed from the EIC. Since the $S(CC_{INI})$ value can only tell us if there is at least one true peak in this EIC, *TIPick* will go back to the peak enhancement step 1 and then proceed the same procedure to recalculate the $S(CC_{INI})$ value in the same EIC after omitting the largest peak.

The detailed workflow of the peak picking process of *TIPick* is illustrated in Fig. 1b. For the iterative peak picking steps for a single standard in our study, please see the supplementary material.

Performance evaluation of TIPick

To evaluate the performance of *TIPick*, we measured the sensitivity (recall), precision, and F-score. The recall is the most common measurement for the evaluation of peak picking algorithms.^[21,22] The recall of the peak picking algorithm is defined as $\frac{TP}{TP+FN}$ where TP is the true positive (the number of true peaks reported by the algorithm) and FN is the false negative (the number of true peaks

not detected by the algorithm). True peaks were experimentally determined. Greater recall values indicate that more true peaks are detected. To evaluate type I errors (false positive) of *TIPick*, the precision value is used and defined as $\frac{TP}{TP+FP}$, where FP is the false positive (the number of peaks reported by the algorithm that are not true peaks). A reported peak should be present in at least two of three duplicate injections. High precision values indicate that more true positives than false positives are detected. A single value, the F-score,^[23,24] combines the precision and the recall and is defined as $\frac{2 \cdot R \cdot P}{R+P}$, where R is the recall and P is the precision.^[25] A perfect peak picking algorithm has an F-score of 1 (100%).

Experimental section

Chemicals

For the NTU MetaCore metabolomics chemical standards library, standards were purchased from Sigma-Aldrich (St. Louis, MO, USA). Forensic drug standards were purchased from Cerilliant Corp. (Round Rock, TX, USA). Mass spectrometry-grade acetonitrile (ACN) was purchased from J.T. Baker (Phillipsburg, NJ, USA), and mass spectrometry-grade water and methanol (MeOH) were purchased from Shalau (Mas d'en Cisa, Barcelona, Spain). Acetic acid was purchased from Merck (Darmstadt, Germany).

Sample preparation

For the analysis of the NTU MetaCore metabolomics chemical standards library, standards were prepared in organic solvents at a concentration of 500–2000 ng ml⁻¹. For forensic drug analysis, a sufficient amount of the standard solutions was added to urine to give spiked urine samples. Spiked urine samples (100 µl) were diluted with 400 µl of deionized water and centrifuged at 15 000 g for 10 min. The supernatant (200 µl) was then subjected to LC-MS analysis.

Chromatographic and mass spectrometric analysis

All standard mixtures and forensic drugs were analyzed with an Agilent 1290 U-HPLC system with a 6540-QTOF (Agilent Technologies, Santa Clara, CA, USA). The mass resolution and the mass accuracies used in this study was 20 000 and 10 ppm, respectively.

For the NTU MetaCore metabolomics chemical standards library analysis, two microliters of the standard mixture was injected into an ACQUITY UPLC HSS T3 column (2.1 mm × 100 mm, 1.8 µm) (Waters, Milford, MA, USA). The mobile phase was composed of 0.1% formic acid in water (solvent A) and ACN (solvent B). The gradient profile was the following: 0–1.5 min: 2% B; 1.5–9 min: linear gradient from 2 to 50% B; 9–14 min: linear gradient from 50 to 95% B; 14–15 min: 95% B. The flow rate was 0.3 ml min⁻¹. For sample ionization, a Jet Stream electrospray ionization source operated with a capillary voltage of 4 kV in positive mode and 3.5 kV in negative mode. The mass scan range was $m/z = 50$ –1700.

For forensic drug analysis, five microliters of sample was injected into an Agilent Poroshell EC-C18 column (2.1 mm × 100 mm, 2.7 µm). The mobile phase was composed of 0.1% acetic acid in water (solvent A) and MeOH (solvent B). The gradient profile was the following: 0–1 min: 2% B; 1–10 min: linear gradient from 2 to 50% B; 10–15 min: linear gradient from 50 to 90% B;

15–17 min: 90% B. For sample ionization, a Jet Stream electrospray ionization source was operated with a capillary voltage of 3 kV in positive mode. The flow rate was 0.4 ml min⁻¹. The mass scan range was m/z = 50–950.

Results and discussion

Reduction of chemical noise by background subtraction

The background subtraction of the *TIPick* algorithm can remove false positive signals. To construct a compound library for metabolomics and toxicological analysis, large quantities of standards are generally analyzed to obtain RT information. False identification of ion signals results in incorrect RT information; *TIPick* can improve the accuracy of the RT information by background subtraction.

The optimal parameter selection greatly affects the performance of *TIPick*. To find the optimal parameters for background subtraction, nine parameter combinations were tested, including different RT windows of 6, 12 or 18 s and mass tolerance (δ) of 5, 10 or 15 ppm relative mass accuracy. The δ value is a user-defined mass tolerance for matching background ions for background subtraction with *TIPick*. A standard mixture of 20 chemicals was used to optimize the background subtraction parameters (Table S-1). The results of the background subtraction with different parameters are presented in Fig. S-1. The baselines of the TICs were high when the δ value was 5 ppm (Fig. 2 a–c); the baselines of the denoised TICs were considerably lower when $\delta \geq 10$ ppm (Fig. 2 d–i). In contrast, scatter plots of RT versus m/z showed that a higher tolerance of both the m/z and the RT shifts resulted in cleaner spectra; however, the probability of incorrectly removing ions increased (see Fig. S-1 for a detailed discussion). Therefore, a δ value of 10 ppm was selected for background subtraction with

TIPick. RT shifts were tested between 6 and 18 s. The background noise of the denoised TICs was still high in the late elution region with an RT window of 6 s (Fig. 2a, d and g), whereas the background noise of the denoised TICs with an RT window of 12 s (Fig. 2b, e, h) was similar to the results obtained with an RT window of 18 s (Fig. 2c, f, i). Thus, an RT window of 12 s was selected for background subtraction with *TIPick*.

To better demonstrate the effectiveness of *TIPick* for removing false positive signals, we selected 3-hydroxyanthranilate (m/z 152.0353 [M-H]⁻) and tiglic acid (m/z 101.0597 [M+H]⁺) as demo target molecules. The EICs of the 3-hydroxyanthranilate standard solution and the blank sample are displayed in Figs. 3a and b. The EICs of the 3-hydroxyanthranilate standard solution showed two signals with RTs of 4.5 and 9 min. Based on the peak height, one might identify the largest peak at 9 min as 3-hydroxyanthranilate. However, this peak is also present in the EIC of the blank sample (Fig. 3b). After background subtraction with *TIPick*, the peak at 9 min disappeared (Fig. 3c). Therefore, the true peak corresponding to 3-hydroxyanthranilate is the signal at 4.5 min. Furthermore, Fig. 3d–f illustrates the effective minimization of false positive signals with *TIPick*. The EICs for the tiglic acid sample (m/z 101.0597) indicated that a peak with an RT of 12 min (Fig. 3d) also appeared in the blank (Fig. 3e). This peak disappeared after background subtraction (Fig. 3f). Therefore, tiglic acid could not be detected in this LC-MS analysis. Falsely identified ion signals generate incorrect RT information and constitute a major problem for the construction of standard libraries. Another example of how *TIPick* significantly reduced the chemical noise to differentiate signals of the standards from those in the solvent blanks is illustrated in Fig. S-2. In summary, the results indicated that the *TIPick* algorithm effectively improved the accuracy of the RT information by performing a background subtraction.

Figure 2. The reduction in the noise level after background subtraction with *TIPick* for different parameters. Total ion chromatograms (TICs) of standard mixtures after background subtraction with the following parameter settings: (a) retention time (RT) window = 6 s, δ = 5 ppm; (b) RT window = 12 s, δ = 5 ppm; (c) RT window = 18 s, δ = 5 ppm; (d) RT window = 6 s, δ = 10 ppm; (e) RT window = 12 s, δ = 10 ppm; (f) RT window = 18 s, δ = 10 ppm; (g) RT window = 6 s, δ = 15 ppm; (h) RT window = 12 s, δ = 15 ppm; (i) RT window = 18 s, δ = 15 ppm.

Figure 3. Extracted ion chromatograms (EICs) of 3-hydroxyanthranilate (a–c) and tiglic acid (d–f): (a) EIC (m/z 152.0353) of the 3-hydroxyanthranilate standard solution before background subtraction; (b) EIC (m/z 152.0353) of the solvent blank; (c) EIC of the 3-hydroxyanthranilate standard solution after background subtraction; (d) EIC (m/z 101.0597) of the tiglic acid standard solution before background subtraction; (e) EIC (m/z 101.0597) of the solvent blank; (f) EIC of the tiglic acid standard solution after background subtraction.

Reducing random noise, enhancing peak intensity and peak picking

TIPick removes chemical noise by background subtraction. However, a mass spectrum also generally contains random noise, e.g. instrument noise. To reduce random noise, *TIPick* extracts chromatograms of target m/z values from LC-MS data of duplicate runs. *TIPick* also performs a peak enhancement by temporarily ignoring data points with low intensities and by averaging duplicate EICs. Because three duplicate injections are generally performed in analytical labs, *TIPick* performs peak enhancements with three duplicates. Also, the intensity of random noise is usually small and because random noise is rarely identical in duplicate experiments, averaging of EICs reduces the random noise and enhances the true peak signal. *TIPick* applies the peak enhancement and peak picking process iteratively to detect peaks.

TIPick uses the $S(CC_{INI})$ value, which is based on the CC_{INI} value, to measure the quality of the chromatograms. A large $S(CC_{INI})$ value indicates the presence of the target chemical. $S(CC_{INI})$ is the ratio of two areas: the denominator of $S(CC_{INI})$ is the minimum area of a right triangle covering CC_{INI} , and the numerator is the area between the hypotenuse of the right triangle and CC_{INI} . In our experience, $S(CC_{INI})$ values of most true peaks are equal to or larger than 0.75 with $W_{avg} = 12$ s. W_{avg} is used as a correction term to avoid the extreme case of a single peak in one EIC ($n = 1$ in the scoring function $\Psi(CC_{INI})$) where only a single data point is present in the CC_{INI} .

We also found that a similar algorithm, Xrea, proposed by Schulz-Trieglaff *et al.*,^[26,27] has been used to filter low-quality proteomics LC-MS spectra. Xrea estimates the quality of tandem mass spectra based on the cumulative normalized intensity (CNI) of a spectrum, which is the cumulative curve of the intensity values after normalization and sorting in increasing order. The major differences between CNI and CC_{INI} are the peak enhancement step and the peak binning step (see 'Peak Picking' in the Theory section). The CNI curves for the noise and the peaks (with different preprocessing parameters) are shown in Fig. 4. The area

under the CNI of the noise (Fig. 4b) was larger than that of the peak (Fig. 4d); however, the difference in score between the noise and the peak was small (<0.5). For the noise, the CC_{INI} and CNI values were identical (Fig. 4b, 4f), showing that the peak enhancement and the peak binning with *TIPick* did not affect the area under the cumulative curve. *TIPick* used peak binning to increase the intensity of the peak and removed the random noise with a user-input intensity factor of 0.1 (Fig. 4g). The resulting area under CC_{INI} of the peak was very small and the $S(CC_{INI})$ value was 0.833 (Fig. 4h), a much larger value than the $S(CC_{INI})$ value of the noise (0.021). CC_{INI} showed a greater difference between the noise and the peaks than CNI. Therefore, CC_{INI} is more effective than CNI for discriminating between the noise and the signal.

We also examined the individual effect of either peak binning or peak enhancement for improving the ability of $S(CC_{INI})$ to discriminate between peaks and noise. The peak binning step enhanced the S/N of the peak (Fig. S-3c) without affecting either the EIC or the CNI of the noise (Fig. S-3a, b). The score of the peak after peak binning was 0.719 (Fig. S-3d), which was significantly higher than the score of the peak without binning preprocessing. The peak enhancement step did not affect either the EIC or the CNI of the noise (Fig. S-3e, f) but removed the random noise in the EIC of the peak (Fig. S-3g). In contrast to the peak binning step, the peak enhancement step did not increase the score value of 0.227 (Fig. S-3h). In this case, both peak binning or peak enhancement steps are required for this peak to be detected; a score larger than the threshold value of 0.75. This suggests that the $S(CC_{INI})$ can be used to detect true peaks more effectively when peak binning and peak enhancement are performed simultaneously. In summary, a higher cumulative curve area indicates the presence of a chromatographic peak (compare Figs. 4d and 4h with Figs. 4b and 4f). For identical noise levels, *TIPick* is able to discriminate the peak from its surrounding noise by providing a higher score to the cumulative curves (compare Figs. 4d and 4h).

TIPick is able to detect small peaks with high background noise in the EIC by applying a peak binning step and a peak enhancement step. Fig. 5a displays a high background noise in the EIC of

Figure 4. Comparison of two different peak picking scoring functions, Xrea (a–d) and *TIPick* (e–h). EICs and the cumulative curves of a peak (4-coumarate, m/z 165.0546) and the noise (m/z 165.0758) selected for this experiment. (a) The EIC of the noise; (b) the cumulative curve of the normalized intensity (CNI) of the noise; (c) the EIC of the peak; (d) the CNI of the peak; (e) the EIC of the noise using *TIPick*; (f) the cumulative curve of the integrated normalized intensity (CC_{INI}) of the peak using *TIPick*; (g) the EIC of the peak using *TIPick*; (h) the CC_{INI} value using *TIPick*.

Figure 5. Examples of the effects of denoising and true peak enhancement with *TIPick*: (a) EICs (m/z 113.0608) of the three duplicate injections of gamma-caprolactone; (b) EICs of (a) after background subtraction; (c) the EIC after background subtraction and peak enhancement.

a gamma-caprolactone standard solution (m/z 113.0608). After background subtraction, the EIC showed considerably less noise (Fig. 5b), but the peak corresponding to gamma-caprolactone could not be detected ($S(CC_{INI}) < 0.75$ for two-thirds of the duplicate EICs) because of the high random noise in the EIC. By averaging three duplicates, the true peak retained its intensity but the random noise was greatly reduced (Fig. 5c). The recalculated $S(CC_{INI})$ value for the enhanced EIC was 0.827, which exceeded the threshold value of 0.75. This example demonstrated that *TIPick* effectively enhanced the signal intensity of a true peak. The power of detecting small peaks therefore improved.

Resolving split peaks

Two major causes for split peaks in our study are instrumental saturation and mathematical processing in the background subtraction algorithm. To resolve the split peak problem, *TIPick* compares the EICs before and after background subtraction and adjusts the mass error tolerance for the EICs.

When *TIPick* detects split peaks, it analyzes the original EIC (before background subtraction) to determine whether the split occurred in the background subtraction process. For example, two peaks in the EIC (m/z 162.1125) of a carnitine standard solution were detected by *TIPick* (Fig. S-4a). However, only one peak existed before background subtraction (Fig. S-4b). The split peak was thus caused by a small peak with the same m/z in the blank solution (Fig. S-4c). When *TIPick* detects split peaks, it compares the number of peaks before and after background subtraction. If only one peak is detected in the original chromatogram, *TIPick* reports the original chromatogram without performing a background subtraction. If the original chromatogram has a similar pattern, *TIPick* will extract an ion chromatogram for the target m/z value with a larger mass error tolerance to determine whether the mass error tolerance has an effect on the peak shape. This action is based on the general observation that the mass accuracy decreases when the detector is saturated with ions. An example of a peak split caused by instrumental saturation is shown in Fig. S-4d. Two peaks in the EIC (m/z 477.2303) of a loperamide standard solution were detected with *TIPick* with a 10 ppm relative mass accuracy setting. By comparing the EICs before and after background subtraction (Fig. S-4d, 4e), *TIPick* concluded that the split peak problem was not a result of background subtraction. *TIPick* therefore moved to the next step, extracting ion chromatograms with a larger mass error tolerance (Fig. S-4f), and the split peak merged to form a single peak after the mass error tolerance was adjusted to 25 ppm. Therefore, a split peak that arises from detector saturation can be recognized by comparing the number of peaks in the EICs obtained for different mass-error tolerance settings.

Application of *TIPick* for toxicological screening and metabolomics

TIPick can be used for toxicological screening and target analysis in metabolomics to detect chemicals in a complicated matrix. To evaluate the performance of *TIPick* for detecting true peaks, we analyzed the 297 chemical mixtures from the NTU MetaCore metabolomics chemical standards library. The following parameters were used for performance evaluation: ppm of relative mass accuracy of 10, winSize of 25, Int_tolerance of 1.5, min_peakwidth of 2, gap_tolerance of 2, SCC_thresh of 0.75, rep_thresh of 2/3 and the split peak recovered tolerance of relative mass

accuracy of 25 ppm. The 297 standards were divided into 26 batches for LC-MS analysis, and data were collected in both the positive and the negative ionization modes. The target m/z values correspond to protonated molecular ions in the positive ionization mode and deprotonated molecular ions in the negative ionization mode. For this analysis, the calculated recall rate, precision and F-score of *TIPick* were 0.99, 0.97 and 0.98, respectively.

To evaluate the performance of *TIPick* for resolving split peaks and how this split peaks resolution step achieved more accurate RT for each peak, the EICs of those 297 standards were inspected with 10 ppm tolerance of relative mass accuracy. There were total of 11 standards considered as split peaks (Table S-3). We manually checked the maximum intensity of a peak to obtain the RT of each standard. This definition could be applied for both symmetric and asymmetric peaks. The true RT in this study was obtained by averaging the RTs of triplicate injections. Table S-4 shows that *TIPick* correctly resolved 10 out of 11 split peaks. Table S-5 shows the detailed results of the 11 split peaks resolved by *TIPick*. Nine out of ten RT differences between reported RT and true RT value were less than 0.17 s. While there are limits to achieving exact comparisons, we find that the reported RT was far more accurate than two other algorithms. The *TIPick* method also resolved more split peaks in this comparison. For details of the accuracy of RT when dealing with split peaks and comparison of the algorithms, please see the supplementary material in Tables S-3 to S-5.

To evaluate the applicability of *TIPick* for toxicological screening, we spiked 51 forensic drugs in human urine (Table S-2). *TIPick* was then used to detect the peaks in LC-MS chromatograms. *TIPick* was able to detect all 51 forensic drugs. Figure 6(a) displays a high background noise in the EIC of buprenorphine (m/z 468.3108). Because of the high background noise, it was difficult to identify buprenorphine by visual inspection. *TIPick* performed background subtraction of a blank sample (urine without a spiked forensic drug) and calculated the $S(CC_{INI})$ value of the EIC. As indicated in Fig. 6(b), the background subtraction provided a cleaner chromatogram with a $S(CC_{INI})$ value of 0.812 (the threshold value was set to 0.75).

TIPick can also be applied for target analysis in metabolomics to identify true peaks in samples. Figure 6c shows the EIC (m/z 137.0244) of a plasma sample obtained from a patient. This peak disappeared after background subtraction with *TIPick* (Fig. 6d), showing that *TIPick* successfully reduced the type I error of compound detection.

In summary, we demonstrated that *TIPick* is an effective algorithm for reducing false positives when identifying target compounds. However, caution should be taken when using *TIPick* for toxicological screening, and a subsequent identification procedure (e.g. RT and mass fragment analysis) should ensue to confirm the presence of the target compound in the sample.

Conclusions

TIPick, a new denoising and peak picking algorithm, was developed for target analysis to detect true peaks in LC-MS chromatograms. *TIPick* performs background subtraction by subtracting ions in the chromatogram of a blank sample and effectively removes false positive signals. By applying peak binning and peak enhancement, *TIPick* detects small peaks in the presence of high background noise. Furthermore, *TIPick* is able to detect tailing peaks and split

Figure 6. Application of the *TIPick* algorithm for toxicological screening and metabolomics: (a) EIC (m/z 468.3108) of the buprenorphine-spiked urine sample; (b) EIC of (a) after background subtraction; (c) EIC (m/z 137.0244) of a plasma extract; (d) EIC of (c) after background subtraction.

peaks by performing peak picking without consideration of the shape of the peak. *TIPick* was successfully applied to construct and analyze the NTU MetaCore metabolomics chemical standards library. The application of *TIPick* for toxicological screening and metabolomics showed that *TIPick* is effective for reducing false positive results. *TIPick* can benefit analytical scientists by automating the discovery of target compounds in complicated sample matrices with a high sensitivity and accuracy.

Acknowledgements

The authors acknowledge I-Lin Tsai for LC-MS experiments, and Chun-Lin Tu, Yeu-Chern Harn, and Jia-Wei Liu for implementing and testing parts of *TIPick* functions. TJH and YFT developed the algorithm. SYW supplied additional analysis requested by the reviewers. CHK and YFT designed the experiments. GYC

performed the LC-MS experiments. This work was financially supported by the Taiwan National Science Council (grant numbers NSC99-2627-B002-018, NSC99-2321-B-002-042 and NSC100-2325-B-002-041). National Taiwan University's Laboratory of Computational Molecular Design and Detection and the Department of Computer Science and Information Engineering contributed resources for this study. We are thankful for the suggestions made by the reviewers

Supporting information

Supplementary electronic material for this paper is available at the Wiley InterScience website.

References

- [1] S. Ojanperä, A. Pelander, M. Pelzing, I. Krebs, E. Vuori, I. Ojanperä, *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1161.
- [2] R. Ramanathan, M. Jemal, S. Ramagiri, Y. Q. Xia, W. G. Humphreys, T. Olah, W. A. Korfmacher, *J. Mass Spectrom.* **2011**, *46*, 595.
- [3] J. L. Campbell, J. C. Le Blanc, *Bioanalysis* **2012**, *4*, 487.
- [4] S. Ojanperä, I. Ojanperä, *LC GC EUROPE* **2005**, *18*, 607.
- [5] H. K. Lee, C. S. Ho, Y. P. H. Lu, P. S. J. Lai, C. C. Shek, Y.-C. Lo, H. B. Klinke, M. Wood, *Anal. Chim. Acta* **2009**, *649*, 80.
- [6] F. Hernández, J. V. Sancho, M. Ibáñez, S. Grimalt, *TrAC, Trends Anal. Chem.* **2008**, *27*, 862.
- [7] C. M. Fleming, B. R. Kowalski, A. Apffel, W. S. Hancock, *J. Chromatogr. A* **1999**, *849*, 71.
- [8] V. P. Andreev, T. Rejtar, H. S. Chen, E. V. Moskovets, A. R. Ivanov, B. L. Karger, *Anal. Chem.* **2003**, *75*, 6314.
- [9] M. Katajamaa, M. Oresic, *J. Chromatogr. A* **2007**, *1158*, 318.
- [10] W. Windig, J. M. Phalp, A. W. Payne, *Anal. Chem.* **1996**, *68*, 3602.
- [11] Y. Li, H. Qu, Y. Cheng, *Anal. Chim. Acta* **2008**, *612*, 19.
- [12] W. Windig, *Chemom. Intell. Lab. Syst.* **2005**, *77*, 206.
- [13] T. Ueno, T. Sueyoshi, Y. Takegami, *J. Mass Spectrom. Soc. Jpn* **1974**, *22*, 95.
- [14] H. Zhang, Y. Yang, *J. Mass Spectrom.* **2008**, *43*, 1181.
- [15] P. Zhu, W. Ding, W. Tong, A. Ghosal, K. Alton, S. Chowdhury, *Rapid Commun. Mass Spectrom.* **2009**, *23*, 1563.
- [16] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, *Anal. Chem.* **2006**, *78*, 779.
- [17] R. Tautenhahn, C. Bottcher, S. Neumann, *BMC Bioinformatics* **2008**, *9*, 504.
- [18] T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic, *BMC Bioinformatics* **2010**, *11*, 395.
- [19] E. Melamud, L. Vastag, J. D. Rabinowitz, *Anal. Chem.* **2010**, *82*, 9818.
- [20] R. Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria, **2011**.
- [21] P. Du, W. A. Kibbe, S. M. Lin, *Bioinformatics* **2006**, *22*, 2059.
- [22] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins, Y. Huang, *Curr. Genomics* **2009**, *10*, 388.
- [23] W. Hersh, A. Cohen, P. Roberts, H. Rekapalli, in *TREC Notebook*, **2006**.
- [24] L. J. Jensen, J. Saric, P. Bork, *Nat. Rev. Genet.* **2006**, *7*, 119.
- [25] C. J. v. Rijsbergen, *Information Retrieval*, second ed, Butterworths: London, **1979**.
- [26] S. Na, E. Paek, *J. Proteome Res.* **2006**, *5*, 3241.
- [27] O. Schulz-Trieglaff, E. Machtejevas, K. Reinert, H. Schluter, J. Thiemann, K. Unger, *BioData Mining* **2009**, *2*, 4.