Plant Biotechnology 29, 109–113 (2012)
DOI: 10.5511/plantbiotechnology.11.1115a

Note

MatchedIonsFinder: A software tool for revising alignment matrices of spectrograms from liquid chromatography-mass spectrometry

Naoki Yamamoto[1,a], Tatsuya Suzuki[2], Takeshi Ara[1], Nozomu Sakurai[1], Sayaka Shinpo[1], Yoshihiko Morishita[1], Ryosuke Sasaki[1], Taneaki Tsugane[2], Hideyuki Suzuki[1], Daisuke Shibata[1,*]

[1] Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan; [2] Chiba Prefectural Agriculture Research Center, Chiba 289-1223, Japan
* E-mail: shibata@kazusa.or.jp Tel: +81-438-52-3947 Fax: +81-438-52-3948

Received October 21, 2011; accepted November 15, 2011 (Edited by T. Demura)

Abstract Several software programs that facilitate processing of data from multiple spectrograms obtained by liquid chromatography-mass spectrometry are available; these programs align features of the ions identified in separate samples into a matrix for further computational analysis. Generally, most alignments within a given matrix are correct, but some alignments seem incorrect, and incorrect alignments must be revised manually. Here, we developed a software tool, MatchedIonsFinder, that revises alignments using a defined algorithm. This program was used to revise an aligned matrix of tomato fruit metabolite ions produced by metabolome analysis software. The matrix revised using the program was comparable to the matrix that was revised manually.

Key words: Comparative metabolome analysis, feature alignment, LC-MS, mass spectrometry.

Metabolomics approaches using liquid chromatography-mass spectrometry (LC-MS) have been widely used in recent plant research (Allwood and Goodacre 2009). A diverse array of secondary metabolites are found in plant species; therefore, LC-MS, which can separate and identify many secondary metabolites, is particularly suited to plant metabolomics research. Metabolome data must be processed before it is analyzed, and metabolome data analysis typically involves multivalent analyses. Typical data processing progresses through multiple stages, including filtering, feature detection, alignment, and normalization; various software tools for metabolome data processing are now available, and many have been reviewed by Katajamaa and Orešič (2007).

Features of detected ions, specifically retention time, mass-to-charge ratio(m/z), and mass fragment pattern, are aligned in a matrix by software. Generally, most alignments in a given matrix are satisfying, but some seem incorrect; consequently, most matrixes require some manual revision of alignments. It is likely that drift in the retention times of particular ions between chromatograms causes these incorrect assignments. For example, the ion peak *Fcd* in chromatogram "c" in the alignment matrix is assigned incorrectly (Figure 1). In this case, the retention time differences among three ions *Fcp*, *Fcd*, and *Fcq* are shorter than those of the corresponding ions *Fap*, *Fab*, and *Faq*, respectively in this region; consequently, *Fab* and *Fcd* are incorrectly identified as different in the matrix. Therefore, when we revise the matrix manually, the drift in retention time is taken into account. To our knowledge, no software tool that revises such aligned matrices is currently available.

Here, we developed a software tool, MatchedIonsFinder, that revises matrices of aligned ion features derived from LC-MS data. To evaluate this tool, we used it to revise a matrix containing processed metabolome data from three tomato cultivars and compared this revised matrix with a manually revised version of the same matrix.

We developed MatchedIonsFinder to revise positional relationships in a matrix of features of ions detected by LC-MS chromatograms; the original positional relationship was generated with existing alignment software. The algorithm encoded in the program is

Abbreviations: LC-MS, liquid chromatography mass spectrometry; m/z, mass-to-charge ratio.

[a] Present address: Hitz Biomass Developing Collaborative Research Laboratory, Graduate School of Engineering, Osaka University, Suita, Osaka 565–0871, Japan

This article can be found at http://www.jspcmb.jp/
Published online February 20, 2012

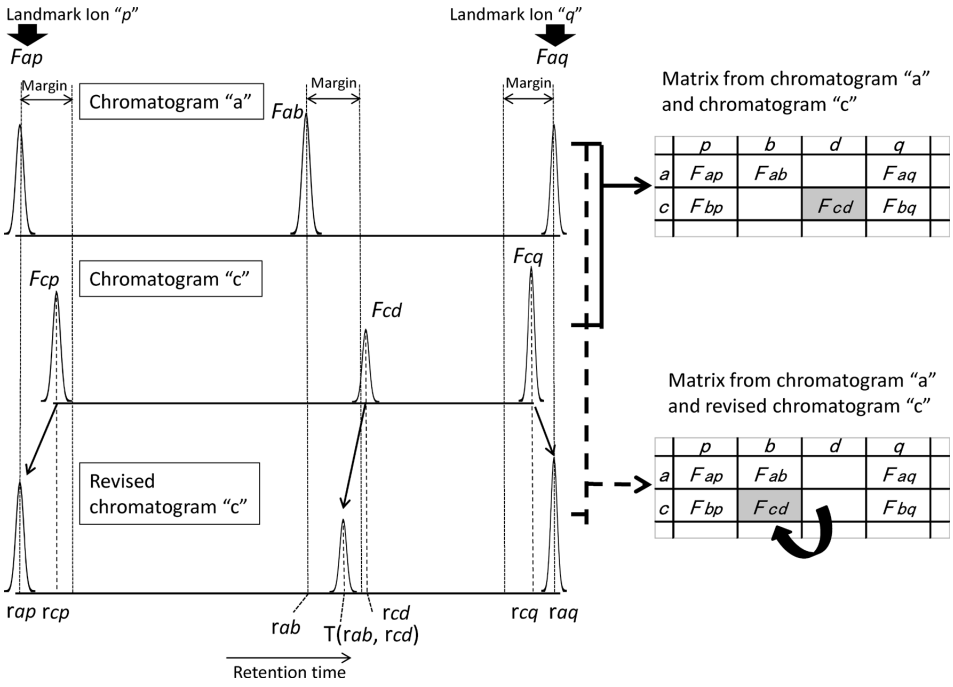

Figure 1. Revision of incorrect assignment of ions in an alignment matrix. A part of the matrix of chromatogram "a" and "c" aligned by alignment software is shown (upper right), as are the original chromatograms (upper and middle left). Landmark ions are *Fap*, *Faq*, *Fcp*, and *Fcq*. The retention times of these ions are shown as rap, r*cp*, r*ab*, r*cd*, r*cq*, and r*aq*. The revised retention time of *Fcd* is T(r*ab*, r*cd*). In this case, *Fcd* is judged to be the same ion as *Fab* within the margin of retention time, and then *Fcd* is aligned in the same column as *Fab* (lower right).

designed to adjust incorrect positioning of ions in the matrix that may be caused by drifts in retention times between each chromatography run. First, the program identifies standard ions to create landmarks, each of which is detected via an ion column in the matrix. The relationships among the landmark ions are not revised at any point in the editing processes. The relationships of ions that migrate between pairs of landmark ions are revised by comparing retention times of the ions among chromatograms (Figure 1).

A description of the processes used in the algorithm is shown in Figure 2. The first landmark is set as a null ion with a retention time of zero. We define an aligned matrix $\boldsymbol{A}$ with ion features of M columns representing the different chromatograms and N rows representing the detected ions.

$$A = \begin{pmatrix} \boldsymbol{F}_{10} & \boldsymbol{F}_{20} & \cdots & \boldsymbol{F}_{\mathrm{M0}} \\ \boldsymbol{F}_{11} & \boldsymbol{F}_{21} & \cdots & \boldsymbol{F}_{\mathrm{M1}} \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{F}_{1\mathrm{N}} & \boldsymbol{F}_{2\mathrm{N}} & \cdots & \boldsymbol{F}_{\mathrm{MN}} \end{pmatrix}$$

$\boldsymbol{F}_{mn}$, which is the element of $\boldsymbol{A}$ on row "n" in a chromatogram "m", is denoted as follows. $\boldsymbol{F}_{mn}$ = [d_{mn}, r_{mn}, z_{mn}, i_{mn}] or [null, null, null, null]. $1 \le m \le M$, $1 \le n \le N$. "d_{mn}", "r_{mn}", "z_{mn}", and "i_{mn}" represent an ion feature identifier, retention time, m/z, and intensity of the ion, respectively, on row "n" in chromatogram "m". As the first landmark, $\boldsymbol{F}_{m0}$=[null, 0, null, null], $1 \le m \le M$].

To relate two independent ion features, $\boldsymbol{F}_{ab}$ and $\boldsymbol{F}_{cd}$ (a, $c \in [1,M]$, b, $d \in [1,N]$), we introduce the value T(r_{ab}, r_{cd}) that is calculated as the revised retention time of $\boldsymbol{F}_{cd}$ by assuming that the retention time r_{ab} of $\boldsymbol{F}_{\mathrm{ab}}$ is correct. The T values for all possible pairs of independent detected features, including reciprocal pairs, are calculated. T(r_{ab}, r_{cd}) is used to judge whether ion d of chromatogram c ($\boldsymbol{F}_{cd}$) should be treated as though it is identical to as ion b of chromatogram "a" ($\boldsymbol{F}_{ab}$) in the revision process. When feature $\boldsymbol{F}_{ab}$ migrates between the closest landmarks, ion p and ion q of chromatogram "a", $1 \le p < b$, $d < q \le m+1$, we define

$$\mathrm{T}(\mathrm{r}_{ab}, \mathrm{r}_{cd}) = \{(\mathrm{r}_{aq} - \mathrm{r}_{ap}) / (\mathrm{r}_{cq} - \mathrm{r}_{cp})\} * (\mathrm{r}_{cd} - \mathrm{r}_{cp}) - \mathrm{r}_{ap}$$

To revise ion features between the landmarks, ion p and ion q, the program starts to find the uppermost feature $\boldsymbol{F}_{\mathrm{i}(p+1)}$ in the column $(p+1)$. Pairwise relationships of other features ($\boldsymbol{F}_{gh}$, $1 < g < M$, $(p+2) \le h < N$) between the landmarks are judged by T($\mathrm{r}_{i(p+1)}$, r_{gh}). If a single feature is found as the same ion with $\boldsymbol{F}_{i(p+1)}$ within the margins of retention time and m/z, the judgment process is performed as shown in Figure 3. If multiple features are found to be the same ion as $\boldsymbol{F}_{i(p+1)}$ in a chromatogram, only one of the features (specifically the feature with the T value closest to $\mathrm{r}_{i(p+1)}$) is chosen. According to the judgment, features are moved or swapped, which results in a new $N \times M$ matrix. Next, the program finds the lower ion feature neighboring $F_{i(p+1)}$ in the $(p+1)$ column in the

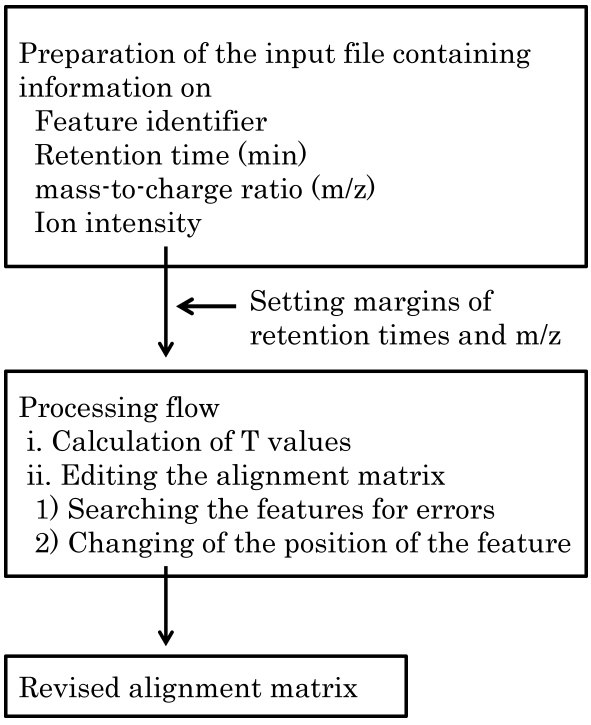

Figure 2. Flow chart of the algorithm of the MatchedIonsFinder calculation. For editing by MatchedIonsFinder_MSMS, the file of information of mass fragmentation spectrum is required (optional).

new matrix. Revisions are repeatedly carried out in this manner. This type of processing is reiterated until the bottom of the $(p+1)$ column is reached and feature finding for ion $(p+1)$ is complete. Feature finding for ion $(p+2)$ proceeds using the same processes. These processes are repeated sequentially until feature finding for ion $(q-1)$ is complete. This entire process is repeated for the ion features between the next pair of landmarks, q and r $(q<r)$, and the process is reiterated until the final landmark, x, is reached. For ion features that appear after the last landmark x,

$$T(r_{ab}, r_{cd}) = r_{cd} + r_{cx} - r_{ax}$$

The revision is carried out successively as shown for ions between landmarks, and finally the $N \times M$ matrix $\boldsymbol{A}_x$ is generated.

We developed another software package, MatchedIonsFinder_MSMS, that processes LC-MS chromatogram data that contains information on mass fragmentation patterns. Rather than using landmark ions found all throughout a column for an ion in the matrix, as MatchedIonsFinder does, MatchedIonsFinder_MSMS uses two pairs of landmark ions in two chromatograms. Calculation of T values and revision processes are the same in both programs, except in cases where $\boldsymbol{F}_{ab}$ and $\boldsymbol{F}_{cd}$ are judged to be the same based on retention times; if mass fragmentation patterns are unmatched in these cases, $\boldsymbol{F}_{ab}$ and $\boldsymbol{F}_{cd}$ are treated as distinct ions, by which reliable revision is expected. A tab-delimited text file that includes feature identifiers, m/z values, and intensity of mass fragments is required when using MatchedIonsFinder_MSMS.

Aligned matrices from major alignment programs such as Mass Profiler Professional (Agilent Technologies, Inc.), MZmine (Tomáš et al. 2010), MetAlign (Arjen et al. 2009), or MarkerLynx (Waters Co.) are converted to the *MatchedIonsFinder* format by simple self-made scripts.

To assess the effectiveness of MatchedIonsFinder and MatchedIonsFinder_MSMS, we applied each program to a matrix of aligned features of tomato metabolites. Three tomato cultivars, *Lovely-Ai* (MIKADO KYOWA SEED Co., Ltd., Tokyo, Japan), *House-Momotaro* (Takii and Co., Ltd., Kyoto, Japan), and *Furikoma* (National Institute of Vegetable and Tea Science, Mie, Japan) were planted in a well-controlled field in the agricultural experimental station of the Chiba Prefectural Agriculture Research Center (Chiba, Japan) on June 28th, 2010; matured fruits from these plants were harvested in September or October. *Lovely-Ai* (n=18), *House-Momotaro* (n=9), and *Furikoma* (n=9) fruits were harvested and then immediately frozen in liquid nitrogen. Fruits from each cultivar were separated into three replicate groups per cultivar with the same number of fruits. The triplicate groups in each cultivar were used for metabolite extraction to generate triplicate samples for metabolome analysis using LC-Fourier transform ion cyclotron resonance-MS (LC-FT/ICR-MS) as described by Iijima et al. (2008). LC-FT/ICR-MS analysis of the triplicate samples from each cultivar resulted in nine chromatograms that were analyzed using

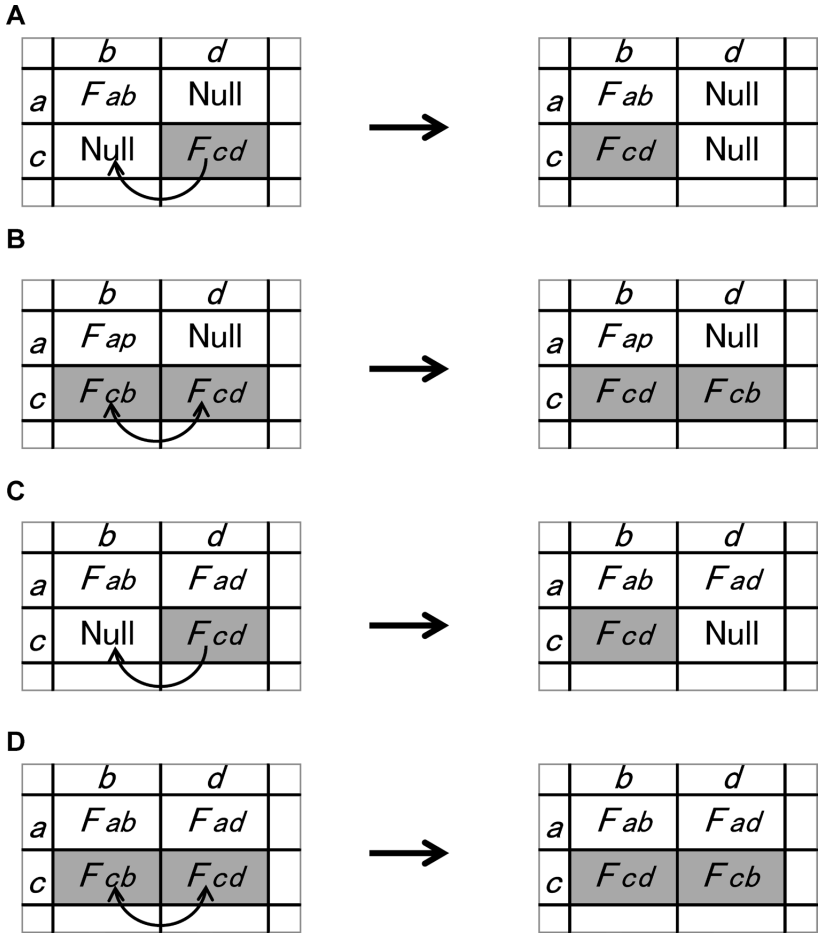

Figure 3. Revision of ion features in an aligned matrix of multiple LC-MS chromatograms. In this figure, $\boldsymbol{F}_{ij}$=[null, null, null, null] is represented as "Null", and an ion feature $\boldsymbol{F}_{ab}$ is assumed to be true. When T(r_{ab}, r_{cd}) represents the most closest value to r_{ab} in chromatogram "*c*", there are four cases for judging which ion features should be aligned with $\boldsymbol{F}_{ab}$ in the same row as shown: (A) $\boldsymbol{F}_{cd}$ is moved into the position of the matrix, row *c* and column *b*. (B) if r_{ab} is closer to T(r_{ab}, r_{cd}) than to T(r_{ab}, r_{cb}), the position of $\boldsymbol{F}_{cd}$ and $\boldsymbol{F}_{cb}$ is swapped; (C) if T(r_{ab}, r_{cd}) is closer to r_{ab} than to r_{cb}, $\boldsymbol{F}_{cd}$ is moved into the position of the matrix, row *c* and column *b*; (D) if r_{ab} is closer to T(r_{ab}, r_{cd}) than to T(r_{ab}, r_{cb}) and, moreover, if and T(r_{ab}, r_{cd}) is closer to r_{ab} than to r_{cb}, the position of $\boldsymbol{F}_{cd}$ and $\boldsymbol{F}_{cb}$ is shifted.

Table 1. Revision of an aligned matrix of tomato metabolites using MatchedIonsFinder. The original matrix was produced by the program IonMatch with the retention time margin of 0.3 min and an m/z margin of 4 ppm. These margins were also used for the calculations in MatchedIonsFinder. See the text for the definition of number of ions.

Number of ions	Original matrix	Processed		Curated manually
		MatchedIonsFinder	MatchedIonsFinder_MSMS	
Three	622	868	736	781
Two	765	825	840	876
One	5730	4847	5213	5073
Zero	10160	8502	9147	8844

the program PowerFT for ion feature detection; the features of nine chromatograms were then aligned using the program IonMatch to produce an aligned matrix file. MatchedIonsFinder, MatchedIonsFinder_MSMS, PowerFT and IonMatch that were made in our laboratory are available from our web site (http://www.kazusa.or.jp/komics/tool_en.html). This matrix was then revised using MatchedIonsFinder or MatchedIonsFinder_MSMS separately.

The new matrices of tomato cultivar metabolites that were revised using MatchedIonsFinder and MatchedIonsFinder_MSMS were compared with the original matrix produced using IonMatch that had been revised manually (Table 1). To evaluate the new matrices, we assume that an ion that is detected in two or three of the triplicate samples from a cultivar is reliable or highly reliable, respectively. In contrast, if an ion is detected in only one of the triplicate samples or is not detected at

all in the triplicate samples it is likely to be noise. After revising the matrix using the MatchedIonsFinder or MatchedIonsFinder_MSMS programs, the numbers of highly reliable and reliable ions increased, and these numbers were comparable to the numbers of ions in the matrix that was revised manually. These results indicate that both programs accurately revised the original matrix of ion features that were detected in the LC-MS chromatograms.

MatchedIonsFinder and MatchedIonsFinder_MSMS are written in the Perl language, and the aligned matrices of tomato metabolites shown in this study are available on our web site (http://www.kazusa.or.jp/komics/software/MatchedIonsFinder/index.html). The raw LC-MS chromatograms used in this study are also available for free in the metabolome database MassBase (http://webs2.kazusa.or.jp/massbase/); the accession numbers are MDLC1_25527-25534, 25539, 25546-25553 and MDLC1_25559.

Acknowledgements

This work was supported by a "Research and Development Projects for Application in Promoting New Policies Agriculture, Forestry and Fisheries" from the Ministry of Agriculture, Forestry and Fisheries of Japan, and the Kazusa DNA Research Institute Foundation.

References

Allwood JW, Goodacre R (2010) An Introduction to Liquid Chromatography-Mass Spectrometry Instrumentation Applied in Plant Metabolomic Analyses. *Phytochem Anal* 21: 33–47

Lommen A (2009) MetAlign: Interface-driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Anal Chem* 81: 3079–3086

Iijima Y, Nakamura Y, Ogata Y, Tanaka K, Sakurai N, Suda K, Suzuki T, Suzuki H, Okazaki K, Kitayama M, Kanaya S, Aoki K, Shibata D (2008) Metabolite annotations based on the integration of mass spectral information. *Plant J* 54: 949–962

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

Pluskal T, Castillo S, Villar-Briones A, Orešič M (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11: 395