preparation of individual standard solutions and the CE-TOFMS condition and instruments were as described elsewhere (Hirayama et al. 2009). We prepared 304 standard metabolites for cation datasets. The concentration of all standard metabolites was 50 μM and 200 μM of methionine sulfone was added as an internal standard. Each mixture was separated into four containers, and then three selected metabolites were additionally spiked into the three bottles at different levels to increase their concentration by 15, 30 and 50%. The selected cationic metabolites were N-α-benzenolarginine ethylester, 2,4-dimethylaniline, and S-(5′-Adenosyl)-L-homocysteine (SAH) and were selected based on their different detection sensitivity. For SAH, the divalent ion peaks were used for the following benchmark experiments. Three replicates of all samples were measured on the same instrument on the same day.

The biological test datasets used for other validations originate from previous studies (Soga et al. 2006). We used serum samples from control mice and mice treated with acetaminophen for 2 h prior to analysis. All numerical experiments were conducted on Windows XP x64 with a Xeon 3 GHz CPU and 8 GB memory.

3 Results and discussion

3.1 File converter

To reduce the file size to be generated, the lowest *m/z* values common to all time-points are memorized and only the difference in the adjacent *m/z* values is stored. The actual *m/z* values for all datapoints are then reconstituted using the sum of the lowest *m/z* and their respective differences. In addition, all data stored in *ciff* files are wrapped in a zlib library (http://www.zlib.net/) to further compress the file size. Details on the file format are available from the *JDAMP* website (http://software.iab.keio.ac.jp/jdamp).

Under our routine measurement conditions (Soga et al. 2006), for each CE–MS run, Analyst QS stores raw data in approximately 100 MB for the cation mode and in approximately 150 MB for the anion mode. Analyst QS can export the raw data to CSV or NetCDF. However, this conversion, without any masking of low abundance intensities, results in an approximately 10-fold increase in file size to approximately 1.0–1.5 GB in the CSV, NetCDF and mzXML formats. On the other hand, the *JDAMP* converter produces *ciff* files that are approximately only 120 and 180 MB for cation and anion data, respectively, which can be easily imported into *JDAMP*. Compared with the use of CSV, NetCDF or mzXML files, the file conversion time is also reduced from 20–40 to 3–4 min, on average. These significant improvements contribute to reduce the processing time for subsequent analysis because file-access time, an important variable in processing numerous large data files, is shortened. Common file conversion tools, such as mzStar (http://tools.proteomecenter.org/mzStar.php), Analyst QS and MassHunter, include an option to eliminate signals below a user-defined intensity threshold to prevent this enlargement of outputs. However, such data reduction should not be implemented during file conversion because this might reduce the possibility of finding significance related to the small peaks; therefore, such functions should be implemented in subsequent analytical processes to enable the users to use small data files without additional file conversion.

The *ciff* file contains data indexes to separate data along the mass spectral and electrophoretic axes, and to reduce the access time to a specific data block in the *ciff* file. Compared with common data formats that represent a series of mass spectra, as in CSV or TXT format, which only allow fast data access to mass spectral data, the *ciff* format enables rapid access to both mass and time dimensions, which significantly reduces calculation times for handling electropherograms.

3.2 Software features

Screenshots of the graphical user interface are shown in Figs. 2, 3, and Supplementary Information Fig. S1. First, the data files (converted with *dotMZ*) for two or more groups to be compared are imported (Fig. S1A). Then, the user specifies the options for preprocessing such as a threshold for the signal/noise ratio. The baseline correction with primary and secondary binning is then executed. Spike noise, defined as signals that are continuous in time for less than the user-defined threshold, is also eliminated at this step (Fig. S1B). In the next step, the user can specify criteria for peak selection and select the DP parameters to be used for migration-time normalization (Fig. S1C); these include the distribution of representative peaks over time or the *m/z* axis, and the gap penalties (Baran et al. 2006). After the migration-time alignment is completed (Figs. S1D and S1E), the internal standard(s), commonly used in CE–MS systems to compensate for changes such as ionization efficiency, injection volume and sensitivity of MS (Ohnesorge et al. 2005), must be chosen to normalize the signal intensities to account for systematic bias between separate measurements and to limit variation to biologically significant variation. However, this step can be omitted if not necessary. The detected differences are visualized directly on 2D density plots (time and *m/z* dimension in Fig. 2A). As recently demonstrated (Erny and Cifuentes 2007), such 2D maps of CE–MS data facilitate intuitive visual inspection of large datasets, which enables the identification of relevant redundant ions such as

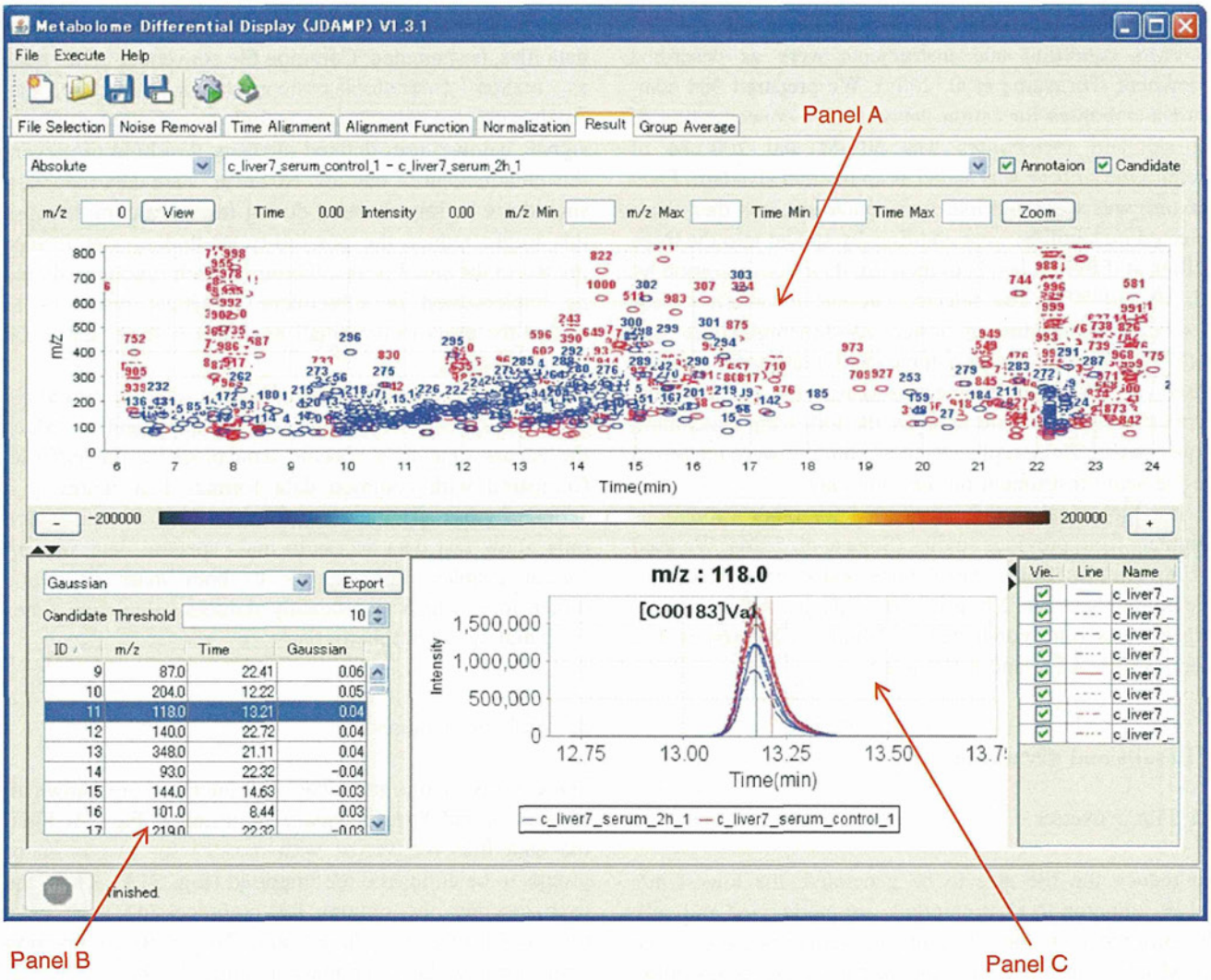


Fig. 2 Screenshots of *JDAMP* results windows. Panel A displays the location of detected significant differences (*red* labels) and of known compounds (*blue* labels). Details of the differences identified are shown in Panel B. An electropherogram overlay is shown for the selected features in Panel C. Other windows, e.g., 2D plots to visualize the averaged intensities within a group and electropherograms of normalized internal standards, are accessible when the respective tab is clicked and the setup window for each process is spawned from the menu or gear icons

fragment ions and adducts, and to differentiate between multiple samples. The map also allows quick overall evaluation of run quality, which is more comprehensive than the total ion electropherogram alone, and yields more readily interpretable information. For example, we empirically know that our CE–MS data always include peaks derived from salts and neutral molecules that appear as vertical smear lines during the first few and last minutes of measurements, respectively. Because of their peak-like appearance, they are not completely removed by baseline correction and the noise-filtering process; however, they are clearly visualized on 2D maps. Such peaks should be eliminated when performing differential analysis using CE–MS data by selecting the corresponding migration time windows for data removal.

To aid visual confirmation of automatically detected differences, a list of significant differences and the corresponding overlaid electropherograms are displayed and linked to each other for easy access to the datapoints of interest (Fig. 2B, C). A user-supplied list of known compounds (chemical standards) can be used to annotate the data and can be visualized on the same figure to facilitate the identification of metabolites in the dataset (Fig. 2A, C), even though further confirmation, such as spiking experiments may be required for reliable identification. *JDAMP* generates structured summary reports, including the detected difference matrix, and corrected electropherograms for whole datasets and a list of detected individual differences for further external analysis with other tools.

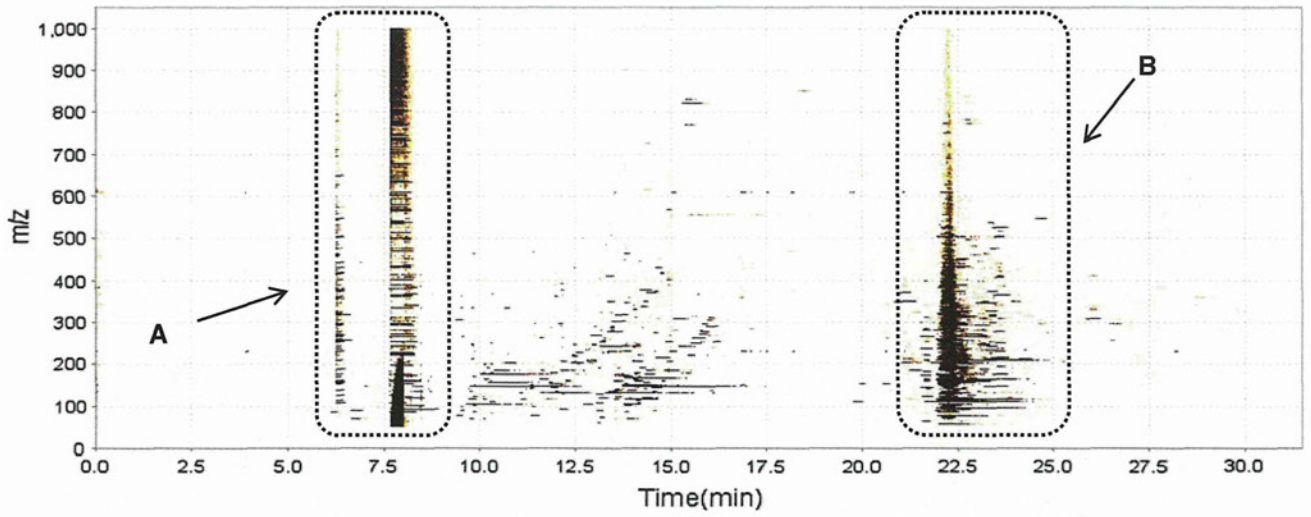


Fig. 3 A typical 2D plot of CE–MS data (time and *m/z* axis) generated after background subtraction and noise filtering. (A) Double vertical smears originating from the early-eluting salt ions or from a sharp baseline drift often occur just after the elution of salt ions. (B) A wider vertical smear derived from a cohort of late-eluting neutral peaks

As described by others (Robinson et al. 2007), the MathDAMP alignment procedure for migration times has some limitations when the datasets are highly dissimilar and users must tune the alignment options to accommodate datasets. As an alternative, we devised the GUI to facilitate prompt quality confirmation by including parameters for alignment algorithms and the range for eliminating unnecessary/undesirable data, and to execute the process iteratively. The optimization options or parameters of the alignment procedure are described in Supplementary Information Text S1 with an example of processing results (Fig. 3).

3.3 Preprocessing for noise reduction

In the preprocessing step, we used a single region of the electropherogram to calculate the noise value, which was used as a threshold to remove noise of low intensity. Supplementary Information Fig. S2 shows the total ion electropherograms and extracted electropherograms of mouse serum datasets. Except for the region around the peaks derived from the analytes and neutral molecules, the deviations are almost constant, and noise was clearly removed. When *JDAMP* is applied to non-CE–MS systems, the current denoising method may not completely eliminate all of the noise across the chromatogram because in LC–MS, for example, such noise generally changes due to a variable mobile phase composition (gradient) resulting in more variable background drift and noise levels.

3.4 Alignment of multiple datasets

Figure 4 and Supplementary Information Fig. S3 depict the differences in migration times between matched peaks in two samples before and after the alignment procedure. The average standard deviations of the migration time difference between five comparisons were reduced from 0.260 min (0.64%) to 0.0190 min (0.047%). In the alignment procedure, although the datasets included a few mismatched representative peaks in the DP phase, most of the correctly matched peaks allowed us to optimize the parameters for Eq. 1 and to produce accurate alignments.

Overall, migration alignment is very useful to correctly match the corresponding signals for differential analysis. However, the electric current condition in the capillary during measurement shows different profiles and is a possible factor that affects the migration time shift and therefore the quality of the alignment results, (Supplementary Information Fig. S4). Variation in the pH of the formic acid solution is also a possible factor responsible for migration time variation. Although the peaks with faster electrophoretic mobility were correctly aligned, the peaks derived from neutral molecules migrating after 22.5 min (Fig. 4A) showed greater variance and were not accurately aligned (Fig. 4B). These peaks represent the main source of poorly aligned signals. However, this part of the data should be discarded or should not be used in subsequent processing because the separation is non-electrophoretic and this part of the data represents neutral molecules.

3.5 Differential detection performance

Methods for peak detection and deconvolution for LC–MS and GC–MS have been developed (Halket et al. 1999; Vivo-Truyols et al. 2005a, b). Although a similar method for CE peaks has been proposed (Garcia-Alvarez-Coque et al. 2005; Wee et al. 2008), its application to actual data

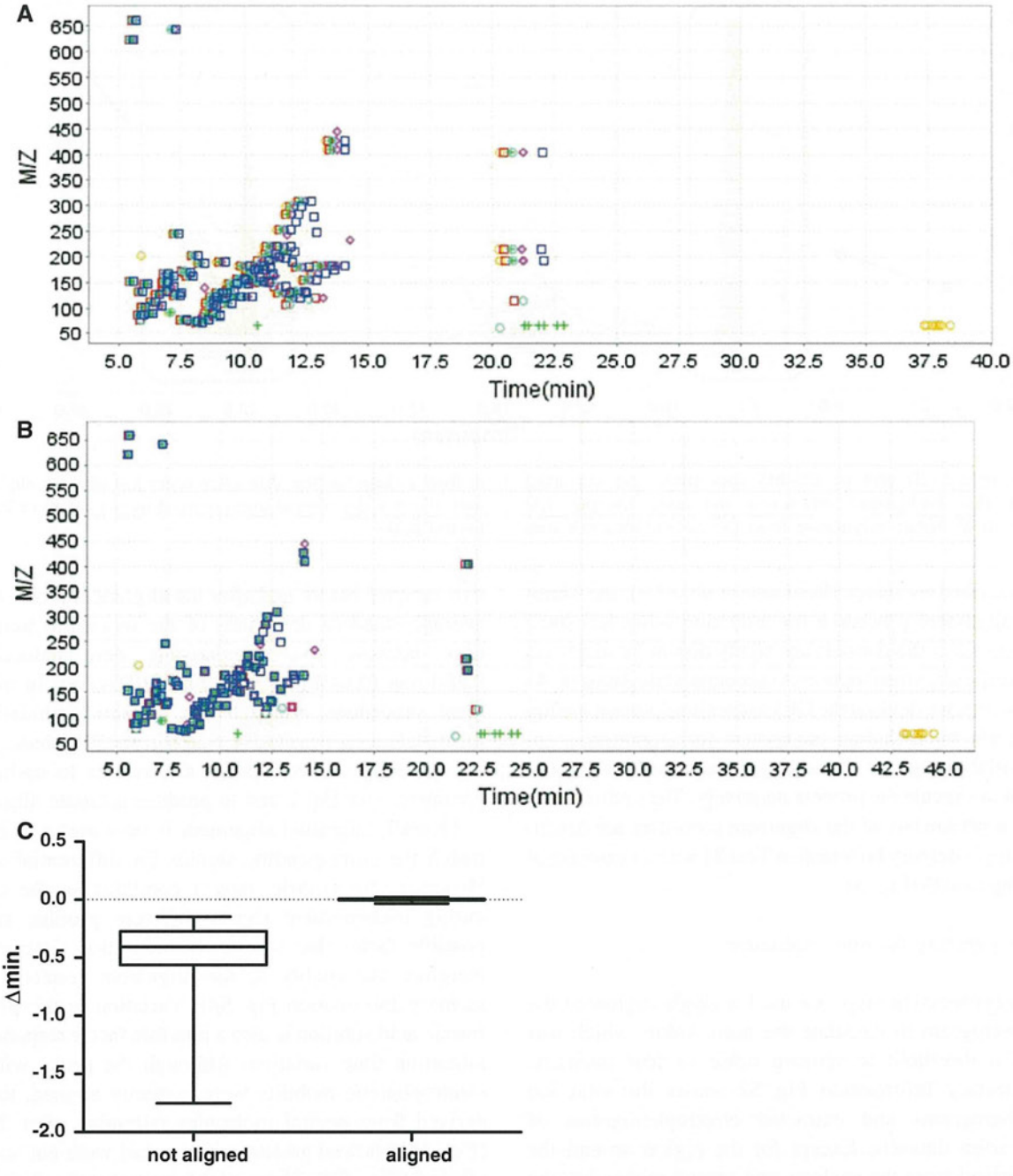


Fig. 4 Migration time alignment results using a standard metabolite mixture. 2D plots (migration time and *m/z* axis) of (**A**) and (**B**) shows actual and normalized representative peak locations, respectively. A total of six samples were aligned simultaneously in this case, and the representative peaks derived from a sample are colored with a single color. Box-and-whisker plots show the difference in migration times of the same representative peaks between two samples (Y-axis, Δmin) in (**C**). The horizontal lines in the box indicate the first quartile, median, third quartile, and the whiskers indicate the maximum and minimum values. Orphan peaks that did not have a matching peak in the corresponding sample and the misaligned peaks in DP phases were eliminated. Plots for other sample combinations and plots showing all differences without elimination of the unmatched peak data are shown in Supplementary Information Figure S3

requires smoothing for noise reduction (Liu et al. 2003; Vivo-Truyols et al. 2005a, b), a process that remains controversial because smoothing distorts the peak area (Wallace et al. 2004). MathDAMP uses the Douglas–Peucker algorithm (Wallace et al. 2004) to select peaks, but only for migration-time alignment, and avoids peak area-based differential feature identification to bypass CE–MS peak detection difficulties. To evaluate the two approaches, we implemented a peak area-based method for difference detection (named area-based detection) and compared its

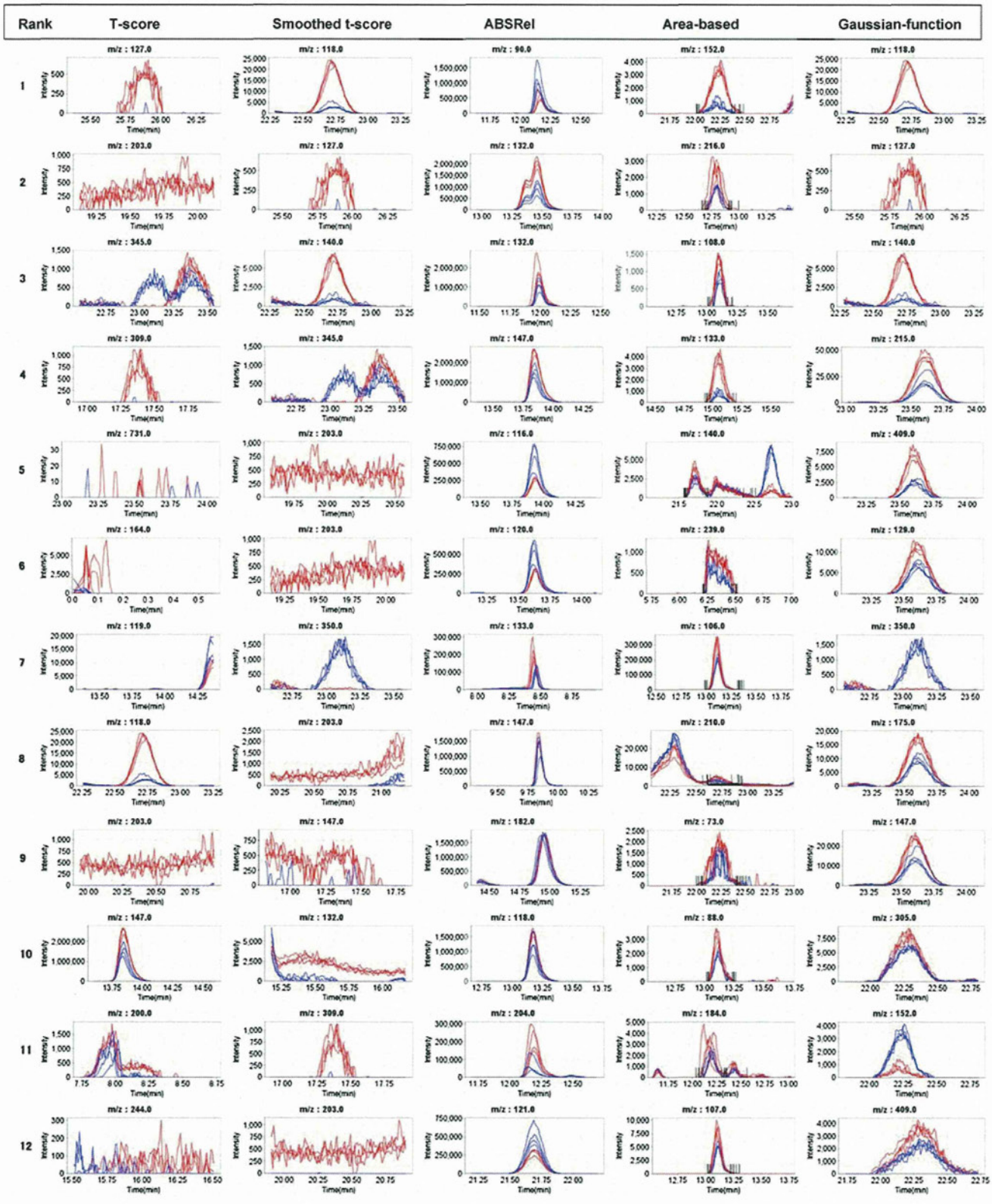


Fig. 5 Overlaid electropherograms of the results ranked in the top 12 from calculations performed using datapoint-by-datapoint *t*-score, smoothed *t*-score, ABSRel, area or Gaussian area functions. The *red* and *blue* curves represent the peaks for the samples and control datasets, respectively

performance with the datapoint-by-datapoint methods. The criteria for the latter included ABSRel, moving average *t*-score using the selected datapoints and the four preceding and subsequent datapoints in the time dimension (named smoothed *t*-score), and Eq. 1 (named Gaussian function).

The ability of *JDAMP* to detect differences was tested using a standard mixture and the results are summarized in Supplementary Information Table S1. Overall the Gaussian function best ranked the N-a-benzenol arginine ethylester metabolites whose concentration was increased compared with the other detection criteria, while the *t*-score showed the worst performance. For example, N-a-benzenol arginine ethylester, which showed high detection sensitivity, was ranked first in the 30 and 50% differentiated solutions and third in the 15% differentiated solution. By contrast, the divalent ion of SAH, which showed low detection sensitivity, was differentially selected only when spiked at an additional level of 30 and 50% and was not found in the 15% spiked samples among 1000 signal rankings based on the *t*-score and Gaussian criteria. For the Gaussian-based results with 2,4-dimethylaniline, even though the accuracy was greater than with the smoothed *t*-score, the rapid deterioration of the results with decreasing spiked amounts suggest that the Gaussian method did not improve the accuracy of peak detection for small peaks or smaller differences. In the datasets used for these validation experiments, a relatively high baseline (background noise), possibly due to lock mass errors or related phenomena, was observed and incomplete elimination of the background yielded a large number of false-positives, which contributed to the deterioration of the differential ranking of 2,4-dimethylaniline and divalent SAH.

For *JDAMP* analysis results using biological samples, Fig. 5 depicts the overlaid electropherograms that were ranked in the top 12 based on these criteria. The overlaid electropherograms for all features ranked within the top 13–50 are listed in Supplementary Information Fig. S5. To reduce false-positive results in the area-based method, peaks that were only found in a few samples across the datasets were eliminated. Here, we used three samples (i.e., 75% of samples in a group of 4 contain the peak) as the threshold and the missing values were set to 0.

Overall, the *t*-score- or smoothed *t*-score datapoint-by-datapoint-based algorithms can detect discriminating peaks when most of the peaks in a group are clearly higher or lower than the peaks in the other groups. However, the electropherogram at 203 *m/z*, ranked 2nd and 9th by the *t*-score method and 5th, 6th, 8th and 12th by the smoothed *t*-score method, showed no clear peaks and the features were scored as significant because of baseline levels that were reproducible between replicates but very different between the two groups. Such false-positives can be rejected by visual inspection of the confirmation plots, demonstrating the importance of this feature. On the other hand, these tests can detect small but clear differences, such as the results at 127 *m/z* and 309 *m/z*, which were ranked 1st and 4th by the *t*-score method and 2nd and 11th by the smoothed *t*-score methods, but were not apparent in the area-based method. This is an important feature of *t*-score-based methods that can be missed (false-negative) by other procedures. Compared with the *t*-score-based method, the results ranked as most significant by the ABSRel index include mainly clear, smooth and high-intensity peaks, even though the algorithm evaluates the datapoints without actual peak detection. Although some peaks manifest significant differences, such as the electropherograms ranked 2nd and 3rd (at 132 *m/z* of $P = 0.017$ and $P = 7.29 \times 10^{-4}$, respectively), the peaks ranked 8th and 9th (147 and 182 *m/z*) exhibited no significant differences ($P = 0.097$ and 0.13, respectively). This is derived from a bias in the ABSRel index, which sometimes highlights signals that are statistically less significant but which show large differences in absolute intensities. The ABSRel index was previously implemented to reduce such bias, which is common when only the absolute difference index is used. However, it cannot be completely eliminated for overwhelmingly large peaks (Baran et al. 2006). By contrast, imperfect alignment or jagged or distorted peaks appear to be responsible for the differences observed in the large internal standard peak at 182 *m/z*, which would be expected to show no difference. Finally, the area-based method could detect peak-like shapes which could be ranked as small, but clearly different peaks (e.g., ranks 1 to 4). However, for effective performance in areas where multiple peaks exist in close proximity, e.g., those ranked 5th and 11th, a more sophisticated peak edge detection algorithm may be needed because some of the peak edges were incorrectly assigned to the neighboring peak and may compromise statistical comparisons.

For the differential analysis of hyphenated MS profiles, both *MZmine* and XCMS perform peak detection, produce lists of statistically significant differences by comparing detected peaks and allow imputation of missing data (Katajamaa and Oresic 2005; Nordstrom et al. 2006; Smith et al. 2006). In addition, a rerun of the integration procedure after dataset alignment to facilitate statistical comparisons is possible because not all of the peaks are detected and aligned in all samples (Katajamaa and Oresic 2005; Nordstrom et al. 2006; Smith et al. 2006). However, the power of their deconvolution algorithms for complex peak shapes and overlapping peaks is unclear. Although datapoint-by-datapoint-based difference detection can bypass such additional procedures, this method alone cannot directly cope with peak deconvolution. However, it can highlight clear differences in irregular and overlapping

peaks, such as the result at 345 *m/z*, which was ranked 3rd by the *t*-score method and 4th by the smoothed *t*-score method. When the objective is to find only statistically significant differences, a low threshold for the peak detection process should be set to allow for the detection of small but significantly different peaks. However, such a procedure involves trade-offs that can compromise either the sensitivity or specificity of the area-based method. Using the Gaussian-based method, the results ranked within the first 12 include signals from both small and large intensities that display, by definition, Gaussian peak-like shapes and also yield small *P*-values. The problem of whether the differences, which are small in absolute terms but statistically significant, represent biologically significant differences needs to be evaluated by further experiments and analyses. Although all methods generate false-positives, Gaussian-based difference-detection appears to minimize their occurrence by combining the high sensitivity of the datapoint-by-datapoint approach and the enhanced specificity of Gaussian fit to normal electrophoretic peaks, thus avoiding noise-related signals. This improvement in accuracy is important to reliably identify discriminating features from large-scale CE–MS datasets. Therefore, the multiple different calculations performed by *JDAMP* represent a major advantage over existing tools and are useful to maximize the detectability of significantly different features.

3.6 Comparison with MathDAMP and MZmine

Using the two criteria, smoothed *t*-score and ABSRel, which are implemented in both MathDAMP and *JDAMP*, the similarities in ranking of differences for the top 50 features in the mouse liver samples are depicted in Supplementary Information Figs. S6A and S6B. Of the detected differences, 64% by ABSRel and 44% by smoothed *t*-score were detected by both tools. ABSRel showed similar profiles to the smoothed *t*-score in MathDAMP and *JDAMP*. Although these differences might predominantly arise from differences in bin borders, the profiles determined using the smoothed *t*-score method were markedly different and were sensitive to the quality of the processing steps prior to the difference detection process. Because the *t*-score method tends to find smaller peaks compared with ABSRel, this discrepancy between MathDAMP and *JDAMP* might explain the differences observed. In the results based on ABSRel, although several peak-shaped results (e.g., a peak at 122 *m/z* (Fig. S6C)) were included only by MathDAMP, a high ranking was assigned to these peaks was due to an overestimation of the significance resulting from incomplete migration time normalization. In the results obtained using the smoothed *t*-score method, those derived from incomplete baseline adjustment, such as in Figs. S6D and S6E, were observed using MathDAMP. Although the former results might be common to both *JDAMP* and MathDAMP and should be eliminated by tuning the options to improve the alignment, the Gaussian-based method implemented in *JDAMP* reduces the detection of the latter cases, as shown in Fig. 5.

With respect to the computation times, the preprocessing takes about 40 to 50 min in both *JDAMP* and MathDAMP because they are based on the same external C++ code module. The migration time alignment process of *JDAMP* requires only a few seconds per dataset while MathDAMP takes about 1 to 2 min under the same conditions. The subsequent steps require 1–2 min for JDAMP and 4–5 min for MathDAMP. Using mouse serum samples (eight datasets), the subsequent procedures including alignment and peak detection took 12 and 38 min for *JDAMP* and MathDAMP, respectively.

We also analyzed the data for the standard metabolite mixture using *MZmine*, a tool that provides peak detection-basis analysis for LC–MS data (Katajamaa et al. 2006). The processing procedure for comparative experiments using *MZmine* is described in Supplementary Information Text S2. Supplementary Information S7 shows typical results obtained using *MZmine*. When the data are converted to mzMXL after eliminating low-intensity signals (<100 cps) to decrease the converted file size, *MZmine* did not detect the expected metabolites and mostly produced false-positives (Supplementary Information Figs. S7A and S7B). In fact, these noise peaks were much larger than other peaks derived from actual metabolite (Figs S7-C and S7-D); therefore, the small deviations in these peaks were, although unexpectedly, detected as differences by *JDAMP* or as peaks by *MZmine*. Only the mzXML data converted without filtering, although each file becomes larger than 1 GB, was successfully used in the subsequent analyses, which might limit the throughput in larger analyses. Using the successfully detected results, alignments with migration time tolerance of 1 and 5% failed to match the peaks even though the average standard deviation of the migration times was 0.64% (Figs. S7E, S7F, and S7G). This result was presumably due to the existence of nearby peaks and, therefore, peak detection with larger peak detection threshold might reduce such instances of misalignment. However, such options will limit the chance of discovery. While the power and utility of *MZmine* for LC–MS data analysis is not questioned here, our results suggest that, at least in its current form, its applicability to the specificities of CE–MS data processing may be limited.

3.7 Advantages and disadvantages of JDAMP

The development of a fully automatic procedure is the ultimate goal to increase throughput for large-scale

metabolomic analysis based on CE–MS data. However, current algorithms optimized for CE–MS data processing such as denoising, peak detection, and migration-time alignment include arbitrary parameters that need to be optimized by the data analysts. To facilitate these tasks, we have developed software tools that feature a simple user interface, improved performance and easier optimization of processing parameters using simple operations with intuitive visual confirmation of the results.

Binning datapoints in the *m/z* domain, as performed by *JDAMP,* results in the loss of high mass resolution obtained by TOF–MS or Fourier Transform Ion Cyclotron Resonance (FT-ICR)–MS, and can limit the identification of adducts, isotopic or fragment-derived peaks. However, while it can considerably facilitate compound identification, the differential detection of features using high-resolution data often requires undesirable or unrealistic computational power and time, and introduces additional steps and hurdles. These include, for example, the need for *m/z* correction across datasets that arise from incomplete *m/z* correction by the MS instrument mass lock feature (Hack and Benner 2002; Soga et al. 2006; Wu and McAllister 2003), an intensity-dependent *m/z* shift due to the signal processing capacity of MS detector (Mihaleva et al. 2008), or peak distortion in the *m/z* dimension (Kempka et al. 2004). For these reasons, we elected to use *m/z* binning as a reasonable trade-off. Once the candidate features are found, the users can easily return to the original high-resolution data using vendor-specific software to extract accurate *m/z* values to facilitate compound identification. In addition, external software should be used to confirm that the observed differences do not originate from different but closely spaced peaks in the *m/z* and migration-time direction, or from corresponding peaks that were assigned to different *m/z* bins due to values near the bin limits.

JDAMP implements metabolite difference detection methods based on both area-based criteria with peak selection and on datapoint-by-datapoint criteria without peak selection. The latter method has significant advantages over peak selection methods for handling irregularly shaped or erroneously missing peaks and can thus enhance the sensitivity of difference detection. Although, empirical mathematical functions to describe electrophoretic peaks have been developed, (Garcia-Alvarez-Coque et al. 2005), the actual peak shapes are, as shown in Figs. 5 or S5, more complicated in biological samples. Multiple factors can influence peak broadening in CE–MS including diffusion, Joule heating, interactions of analytes with the capillary wall, pressure-induced parabolic flow, and negative pressure at the capillary outlet originating from the nebulizing gas (Axen et al. 2007); these can make the peak detection problem more difficult. Although the datapoint-by-datapoint approach is hardly affected by this increased complexity, good results require more accurate migration time normalization than the general approach with peak detection and matching. While most generally used alignment methods to generate matched peak matrix result in other difficulties related to peak splitting or merging (reviewed in Robinson et al. 2007), they require only good peak matching. By contrast, the datapoint-by-datapoint approach requires that the peak maximum is properly matched on the normalized electropherograms, otherwise false-positive signals are often generated. However, easy visualization of the original overlaid electropherogram as implemented in *JDAMP* allows to rapidly exclude these signals.

Because of uncertainty in the number of total features or peaks in the dataset, we did not implement *P*-value corrections such as Bonferroni's correction (Shaffer 1995), which can conservatively correct for multiple hypothesis testing in the *t*-test. Users should be aware that false-positive results will be generated from any such multivariate analyses (more likely for larger *P*-values) and could perform simple correction by estimating the total number of peaks or preferably perform additional experiments to confirm the reproducibility of the original findings. For the same reason that peaks are not used for many calculations, the annotation or elimination of redundant data—arising from isotopic peaks, alternatively charged ions, adducts or fragment ions—is not part of the current *JDAMP* features. However, inspection of the 2D maps can reveal such occurrences as characteristically spaced signals that are vertically well aligned, and allow the user to eliminate these apparently significant but potentially misleading features. In addition, the 2D maps can assist the users to identify and eliminate regions where salt and neutral molecules migrate (visualized as obvious vertical streaks across the datasets). However, further developments are necessary for automatic elimination of those undesirable results using objective criteria. Instrument-specific artifacts previously reported for Orbitrap MS (Brown et al. 2009), such as instrument-dependent and run-to-run difference, were also observed in CE-TOFMS data. For example, unclear but weak vertical lines sometimes appear migrating just prior (left) to the neutral molecule-derived band. These occasionally observed horizontal bands along electropherograms at 92 *m/z*, which are distinct from background ions used for lock mass, may be derived from contamination of the nitrogen gas. Further studies are needed to store these empirical rules and to implement general or ad hoc noise filters.

The *JDAMP* file converter and specific file format provide important benefits, even when handling a relatively small number of datasets and are essential when hundreds of datasets are analyzed on a routine basis to optimize data storage and improve performance. *JDAMP* is a powerful

and rapid tool that identifies significant differences, and is thus useful for initial high-throughput screening of metabolomics datasets. High accuracy *m/z* values to generate compositional formulae and the manual interpretation of mass spectra may be necessary for reliable identification. A number of vendor-supplied software packages, such as Analyst QS, Mass Hunter and Mass Lynx, are user-friendly and are useful for such tasks. However, they lack specific features for automated and reliable differential feature selection between numerous datasets and are thus complementary to *JDAMP*. On the other hand, many other useful tools based on statistical/mathematical software, such as XCMS (Smith et al. 2006), which is based on the R statistical language (University of Auckland; http://www.r-project.org/), MathDAMP (Baran et al. 2006), which is based on Mathematica (Wolfram Research, Inc.; http://www.wolfram.com/), or other recently described software (Allard et al. 2008) based on Matlab (Mathworks, Inc; http://www.mathworks.com/), remain relatively difficult to use, but can offer extra flexibility that is useful for routine analyses or to combine tools with external packages for further analyses. *MZmine* is another powerful tool with the benefit of a sophisticated user interface, but it was developed primarily for LC–MS data analysis (Katajamaa et al. 2006) and, as shown, may be less useful for CE–MS data analysis. The various difference detection methods implemented in *JDAMP* are currently limited to the comparison of two groups, and to evaluate candidate features individually (univariate testing). Pattern recognition technologies, such as support vector machine or partial least square-discriminant analysis, and artificial neural networks, as well as multivariate analyses such as principal components analysis or partial least squares discriminant analysis, have been widely used to simultaneously evaluate multiple peaks and enhance the potential to discriminate between given samples (Acevedo et al. 2007; Mahadevan et al. 2008). To facilitate such multivariable analyses and to enable multiple comparisons between a greater number of groups (>2), *JDAMP* can export intermediate or final results in several formats for downstream use in other software tools. Further development of visual methods for simultaneous comparison of multiple groups is needed (Baran et al. 2007).

JDAMP might be used for instruments other than the ESI-TOFMS used in this study but, for the differential detection approach of metabolic profiles, accurate quantification of signals is a prerequisite to correctly evaluate the significance of the difference. The wider linearity range for quantification in ESI-TOFMS compared with MALDI–MS provides advantages to quantify the difference in biological sources (Ohnesorge et al. 2005). With the use of a supported data converter, *JDAMP* might also be used with data obtained from other types of mass spectrometers, e.g., ion-trap or quadrupole instruments. However, the higher sensitivity of ESI-TOF–MS compared with these techniques (Simo et al. 2008) enhances the limit of detection of small but significant differences with *JDAMP*.

Finally, with the exception of MathDAMP, most of the other currently available software solutions are not optimized for some of the specificities of CE–MS-derived data (peak shape and migration time shifts) and are also based almost exclusively on standard peak detection-based analysis, which offers advantages but also has limitations, as described above. Therefore, rather than replace these tools, *JDAMP* was designed to fill a gap in metabolic data processing and provide an easy-to-use, complementary tool that offers versatile methods to compare metabolite profiles obtained with CE–MS.

4 Concluding remarks

We developed *JDAMP* to offer simplified and faster quantitative differential analysis of high-throughput CE–MS-based metabolomics data. Our software rapidly processes large datasets, detects differences among multiple datasets using different operations, allows visualization of the results using an intuitive and easy-to-use GUI, and can export analysis reports. *JDAMP* enables complementary peak area-based and datapoint-by-datapoint differential feature identification. We expect the software to considerably simplify the analysis of large CE–MS datasets and the identification of discriminatory features such as potential biomarkers. For academic research purposes, the software, manual and animated tutorials are freely available at http://software.iab.keio.ac.jp/jdamp and the source code is available upon request.

Acknowledgments We thank Dr. Yusuke Tanigawara and Dr. Akito Nishimuta of the School of Medicine, Keio University, Dr. Satoshi Yoshida and Dr. Hideki Koizumi of Kirin Holdings, Dr. Akira Oikawa of Riken, and Dr. Eri Shimizu and Dr. Tadahiro Ozawa of Kao Corporation, for valuable discussions. We also thank Maki Sugawara, Hiroko Ueda, Shinobu Abe, and Kazuki Sugisaki of IAB for measurement, data analyses, and programming, and Dr. Ursula Petralia for editing the manuscript. This work was supported by research grants from the Yamagata Prefectural Government and the City of Tsuruoka.

References

Acevedo, F. J., Jimenez, J., Maldonado, S., Dominguez, E., & Narvaez, A. (2007). Classification of wines produced in specific regions by UV-visible spectroscopy combined with support vector machines. *Journal of agricultural and food, 55*, 6842–6849.

Allard, E., Backstrom, D., Danielsson, R., Sjoberg, P. J., & Bergquist, J. (2008). Comparing capillary electrophoresis-mass spectrometry fingerprints of urine samples obtained after intake of coffee, tea, or water. *Analytical chemistry, 80*, 8946–8955.

Axen, J., Axelsson, B. O., Jornten-Karlsson, M., Petersson, P., & Sjoberg, P. J. (2007). An investigation of peak-broadening effects arising when combining CE with MS. *Electrophoresis, 28*, 3207–3213.

Baran, R., Kochi, H., Saito, N., et al. (2006). MathDAMP: A package for differential analysis of metabolite profiles. *BMC Bioinformatics, 7*, 530.

Baran, R., Robert, M., Suematsu, M., Soga, T., & Tomita, M. (2007). Visualization of three-way comparisons of omics data. *BMC Bioinformatics, 8*, 72.

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

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., & Sumner, L. W. (2006). MET-IDEA: Data extraction tool for mass spectrometry-based metabolomics. *Analytical chemistry, 78*, 4334–4341.

Brown, M., Dunn, W. B., Dobson, P., et al. (2009). Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst, 134*, 1322–1332.

Bunk, B., Kucklick, M., Jonas, R., et al. (2006). MetaQuant: A tool for the automatic quantification of GC/MS-based metabolome data. *Bioinformatics, 22*, 2962–2965.

Bylund, D., Danielsson, R., Malmquist, G., & Markides, K. E. (2002). Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *Journal of Chromatography. A, 961*, 237–244.

Erny, G. L., & Cifuentes, A. (2007). Simplified 2-D CE–MS mapping: Analysis of proteolytic digests. *Electrophoresis, 28*, 1335–1344.

Fiehn, O., Kopka, J., Dormann, P., et al. (2000). Metabolite profiling for plant functional genomics. *Nature biotechnology, 18*, 1157–1161.

Fischer, B., Grossmann, J., Roth, V., et al. (2006). Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics, 22*, e132–e140.

Garcia-Alvarez-Coque, M. C., Simo-Alfonso, E. F., Sanchis-Mallols, J. M., & Baeza-Baeza, J. J. (2005). A new mathematical function for describing electrophoretic peaks. *Electrophoresis, 26*, 2076–2085.

Hack, C. A., & Benner, W. H. (2002). A simple algorithm improves mass accuracy to 50–100 ppm for delayed extraction linear matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry, 16*, 1304–1312.

Haimi, P., Uphoff, A., Hermansson, M., & Somerharju, P. (2006). Software tools for analysis of mass spectrometric lipidome data. *Analytical Chemistry, 78*, 8324–8331.

Halket, J. M., Przyborowska, A., Stein, S. E., et al. (1999). Deconvolution gas chromatography/mass spectrometry of urinary organic acids–potential for pattern recognition and automated identification of metabolic disorders. *Rapid Communications in Mass Spectrometry, 13*, 279–284.

Hardy, N. W., & Taylor, C. F. (2007). A roadmap for the establishment of standard data exchange structures for metabolomics. *Metabolomics, 3*, 1573–3890.

Hirayama, A., Kami, K., Sugimoto, M., et al. (2009). Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. *Cancer Research, 69*, 4918–4925.

Karpievitch, Y. V., Hill, E. G., Smolka, A. J., et al. (2007). PrepMS: TOF MS data graphical preprocessing tool. *Bioinformatics, 23*, 264–265.

Katajamaa, M., Miettinen, J., & Oresic, M. (2006). MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics, 22*, 634–636.

Katajamaa, M., & Oresic, M. (2005). Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics, 6*, 179.

Katajamaa, M., & Oresic, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography. A, 1158*, 318–328.

Kempka, M., Sjodahl, J., Bjork, A., & Roeraade, J. (2004). Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry, 18*, 1208–1212.

Lee, R., Ptolemy, A. S., Niewczas, L., & Britz-McKibbin, P. (2007). Integrative metabolomics for characterizing unknown low-abundance metabolites by capillary electrophoresis-mass spectrometry with computer simulations. *Analytical Chemistry, 79*, 403–415.

Liu, B. F., Sera, Y., Matsubara, N., Otsuka, K., & Terabe, S. (2003). Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis. *Electrophoresis, 24*, 3260–3265.

Mahadevan, S., Shah, S. L., Marrie, T. J., & Slupsky, C. M. (2008). Analysis of metabolomic data using support vector machines. *Analytical Chemistry, 80*, 7562–7570.

Mihaleva, V., Vorst, O., Maliepaard, C., et al. (2008). Accurate mass error correction in liquid chromatography time-of-flight mass spectrometry based metabolomics. *Metabolomics, 4*, 171–182.

Monton, M. R., & Soga, T. (2007). Metabolome analysis by capillary electrophoresis-mass spectrometry. *Journal of Chromatography. A, 1168*, 237–246.

Nicholson, J. K., & Wilson, I. D. (2003). Opinion: Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. *Nature Reviews. Drug Discovery, 2*, 668–676.

Nordstrom, A., O'Maille, G., Qin, C., & Siuzdak, G. (2006). Nonlinear data alignment for UPLC–MS and HPLC–MS based metabolomics: Quantitative analysis of endogenous and exogenous metabolites in human serum. *Analytical Chemistry, 78*, 3289–3295.

Ohnesorge, J., Neususs, C., & Watzig, H. (2005). Quantitation in capillary electrophoresis-mass spectrometry. *Electrophoresis, 26*, 3973–3987.

Pedrioli, P. G., Eng, J. K., Hubley, R., et al. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology, 22*, 1459–1466.

Plumb, R., Granger, J., Stumpf, C., et al. (2003). Metabonomic analysis of mouse urine by liquid-chromatography-time of flight mass spectrometry (LC-TOFMS): Detection of strain, diurnal and gender differences. *Analyst, 128*, 819–823.

Reijenga, J. C., Martens, J. H., Giuliani, A., & Chiari, M. (2002). Pherogram normalization in capillary electrophoresis and micellar electrokinetic chromatography analyses in cases of sample matrix-induced migration time shifts. *Journal of Chromatography B, Analytical Technologies in the Biomedical and Life Sciences, 770*, 45–51.

Reo, N. V. (2002). NMR-based metabolomics. *Drug and Chemical Toxicology, 25*, 375–382.

Robinson, M. D., De Souza, D. P., Keen, W. W., et al. (2007). A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC Bioinformatics, 8*, 419.

Ruckstuhl, A. F., Jacobson, M. P., Field, R. W., & Dodd, J. A. (2001). Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer, 68*, 179–193.

Saito, N., Robert, M., Kitamura, S., et al. (2006). Metabolomics approach for enzyme discovery. *Journal of Proteome Research, 5*, 1979–1987.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*, 561–584.

Simo, C., Moreno-Arribas, M. V., & Cifuentes, A. (2008). Ion-trap versus time-of-flight mass spectrometry coupled to capillary electrophoresis to analyze biogenic amines in wine. *Journal of Chromatography. A, 1195*, 150–156.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry, 78*, 779–787.

Soga, T., Baran, R., Suematsu, M., et al. (2006). Differential metabolomics reveals ophthalmic acid as an oxidative stress biomarker indicating hepatic glutathione consumption. *Journal of Biological Chemistry, 281*, 16768–16776.

Soga, T., Ohashi, Y., Ueno, Y., et al. (2003). Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *Journal of Proteome Research, 2*, 488–494.

Styczynski, M. P., Moxley, J. F., Tong, L. V., et al. (2007). Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Analytical Chemistry, 79*, 966–973.

Tautenhahn, R., Bottcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics, 9*, 504.

Vivo-Truyols, G., Torres-Lapasio, J. R., van Nederkassel, A. M., Vander Heyden, Y., & Massart, D. L. (2005a). Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part I: Peak detection. *Journal of Chromatography. A, 1096*, 133–145.

Vivo-Truyols, G., Torres-Lapasio, J. R., van Nederkassel, A. M., Vander Heyden, Y., & Massart, D. L. (2005b). Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part II: Peak model and deconvolution algorithms. *Journal of Chromatography. A, 1096*, 146–155.

Wallace, W. E., Kearsley, A. J., & Guttman, C. M. (2004). An operator-independent approach to mass spectral peak identification and integration. *Analytical Chemistry, 76*, 2446–2452.

Wang, T., Shao, K., Chu, Q., et al. (2009). Automics: An integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics, 10*, 83.

Wee, A., Grayden, D. B., Zhu, Y., Petkovic-Duran, K., & Smith, D. (2008). A continuous wavelet transform algorithm for peak detection. *Electrophoresis, 29*, 4215–4225.

Wittke, S., Fliser, D., Haubitz, M., et al. (2003). Determination of peptides and proteins in human urine with capillary electrophoresis-mass spectrometry, a suitable tool for the establishment of new diagnostic markers. *Journal of Chromatography. A, 1013*, 173–181.

Wong, J. W., Cagney, G., & Cartwright, H. M. (2005). SpecAlign–processing and alignment of mass spectra datasets. *Bioinformatics, 21*, 2088–2090.

Wu, J., & McAllister, H. (2003). Exact mass measurement on an electrospray ionization time-of-flight mass spectrometer: Error distribution and selective averaging. *Journal of Mass Spectrometry, 38*, 1043–1053.

Yoshida, S., Hashimoto, K., Tanaka-Kanai, K., Yoshimoto, H., & Kobayashi, O. (2007). Identification and characterization of amidase-homologous AMI1 genes of bottom-fermenting yeast. *Yeast, 24*, 1075–1084.

Zhao, Q., Stoyanova, R., Du, S., Sajda, P., & Brown, T. R. (2006). HiRes–a tool for comprehensive assessment and interpretation of metabolomic data. *Bioinformatics, 22*, 2562–2564.

Special Feature: Tutorial

Journal of MASS SPECTROMETRY

Received: 18 March 2010 Accepted: 1 June 2010 Published online in Wiley Interscience:

(www.interscience.com) DOI 10.1002/jms.1777

MassBank: a public repository for sharing mass spectral data for life sciences

Hisayuki Horai,[a] Masanori Arita,[a–c†] Shigehiko Kanaya,[d] Yoshito Nihei,[a] Tasuku Ikeda,[a] Kazuhiro Suwa,[b] Yuya Ojima,[a] Kenichi Tanaka,[d] Satoshi Tanaka,[e,f] Ken Aoshima,[e,f] Yoshiya Oda,[e,f] Yuji Kakazu,[a] Miyako Kusano,[c] Takayuki Tohge,[c] Fumio Matsuda,[c] Yuji Sawada,[c,f] Masami Yokota Hirai,[c,f] Hiroki Nakanishi,[f,g] Kazutaka Ikeda,[f,g] Naoshige Akimoto,[h] Takashi Maoka,[i] Hiroki Takahashi,[d] Takeshi Ara,[j] Nozomu Sakurai,[j] Hideyuki Suzuki,[j] Daisuke Shibata,[j] Steffen Neumann,[k] Takashi Iida,[l] Ken Tanaka,[m] Kimito Funatsu,[n] Fumito Matsuura,[o] Tomoyoshi Soga,[a] Ryo Taguchi,[f,g] Kazuki Saito[c] and Takaaki Nishioka[a]*

MassBank is the first public repository of mass spectra of small chemical compounds for life sciences (<3000 Da). The database contains 605 electron-ionization mass spectrometry(EI-MS), 137 fast atom bombardment MS and 9276 electrospray ionization (ESI)-MS^n data of 2337 authentic compounds of metabolites, 11 545 EI-MS and 834 other-MS data of 10 286 volatile natural and synthetic compounds, and 3045 ESI-MS^2 data of 679 synthetic drugs contributed by 16 research groups (January 2010). ESI-MS^2 data were analyzed under nonstandardized, independent experimental conditions. MassBank is a distributed database. Each research group provides data from its own MassBank data servers distributed on the Internet. MassBank users can access either all of the MassBank data or a subset of the data by specifying one or more experimental conditions. In a spectral search to retrieve mass spectra similar to a query mass spectrum, the similarity score is calculated by a weighted cosine correlation in which weighting exponents on peak intensity and the mass-to-charge ratio are optimized to the ESI-MS^2 data. MassBank also provides a merged spectrum for each compound prepared by merging the analyzed ESI-MS^2 data on an identical compound under different collision-induced dissociation conditions. Data merging has significantly improved the precision of the identification of a chemical compound by 21–23% at a similarity score of 0.6. Thus, MassBank is useful for the identification of chemical compounds and the publication of experimental data.

Keywords: MassBank; public database; distributed database; metabolite; spectral similarity

Introduction

Mass spectral data are important experimental data for supporting life science research. Researchers are encouraged to annotate/describe every detail of their experimental data, especially metadata, available to the public at publication of their studies. Full disclosure of supporting experimental data is required for other scientists to confirm the quality of experimental data.[1] However, most mass spectral or supplementary data in journal articles are not fully disclosed because they are published only as figures showing the mass-to-charge ratio (m/z) and the relative intensity values of major peaks.

Although published mass spectral data are valuable research products that should be shared as reference data for the identification of chemical compounds detected by mass spectrometry, their retrieval from journal archives is extremely time consuming. Therefore, mass spectral data as supporting experimental data and as useful research products should be publicly accessible not in figures but in digital format. However, at present there is no public repository for mass spectral data of small chemical compounds except for those of proteomics data. Before considering

* *Correspondence to: Takaaki Nishioka, Institute for Advanced Biosciences, Keio University, 14-1 Banba-cho, Tsuruoka, Yamagata 997-0035, Japan. E-mail: takaaki@sfc.keio.ac.jp*
† *Current address: Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan.*
a *Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0035, Japan*
b *Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan*
c *RIKEN Plant Science Center, Yokohama, Kanagawa 230-0045, Japan*
d *Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan*
e *Biomarkers and Personalized Medicine Core Function Unit, Eisai Product Creation Systems, Eisai Co. Ltd, Tsukuba, Ibaraki 300-2635, Japan*
f *JST, CREST, Kawaguchi, Saitama 332-0012, Japan*
g *Graduate School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan*
h *Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan*
i *Research Institute for Production Development, Kyoto 606-0805, Japan*
j *Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan*
k *Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology, 06120 Halle, Germany*
l *College of Humanities and Sciences, Nihon University, Tokyo 156-8550, Japan*
m *Institute of Natural Medicine, University of Toyama, Toyama 930-0194, Japan*
n *Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan*
o *Faculty of Life Science and Biotechnology, Fukuyama University, Fukuyama, Hiroshima 729-0292, Japan*

the reasons for this, we will briefly review a few currently available mass spectral databases.

Several small-scale databases of mass spectral data of small chemical compounds provide reference mass spectral libraries for metabolite identification. The Golm Metabolome Database (GMD@CSB.DB), established by the Max Planck Institute of Molecular Plant Physiology (Golm, Germany), is a library of GC-MS data of plant metabolites.[2] The METLIN database of the Scripps Research Institute (San Diego, CA, USA) provides 8800 MS^2 data on 1662 metabolites and drugs[3] and the Glycan Mass Spectral Database (GMDB), created by the Research Center for Medical Glycoscience of the National Institute of Advanced Industrial Science and Technology (AIST), Japan, is a library of MS^n data of polysaccharide chains.[4] The Human Metabolome Database (HMDB) of the University of Alberta (Edmonton, Canada) contains liquid chromatography (LC)- and GC-MS data (as PNG images) of 799 and 279 endogenous metabolites reported in the literature that were found in biofluids, respectively.[5] All the electrospray ionization (ESI)-MS^2 data were collected at three different collision energy levels. Two major mass spectral databases, the Mass Spectral Library[6] [the National Institute of Standards and Technology (NIST)/Environmental Protection Agency (EPA)/National Institutes of Health (NIH), USA] and the Spectral Database System (SDBS)[7] of AIST provide 220 000 and 24 000 official mass spectral data, respectively. These national laboratories analyze purified natural and synthetic chemical compounds by electron-ionization mass spectrometry (EI-MS).

In those six databases, all mass spectra were analyzed under fixed, well-controlled experimental conditions. To retain the quality of the data as reference data for the identification of chemical compounds, curators do not mix data in their databases with data analyzed by other research groups.

In the life sciences, different types of mass spectrometers are used to analyze chemical compounds in biological samples because their diverse chemical structure results in different physicochemical properties.[8,9] For example, in most metabolomics studies, GC and LC are coupled to EI-MS and ESI-MS^n, respectively. EI-MS, which applies a standardized analytical method, yields reproducible data for an identical chemical compound. On the other hand, no standard experimental protocol is available for ESI-MS^n. Individual researchers optimized their experimental methods of ESI-MS^n depending on the physicochemical properties of their target chemical compounds. However, slight differences in the experimental methods of ESI-MS^n may yield different mass spectra for an identical chemical compound. Therefore, if a public repository were available, the mass spectral data analyzed by different experimental methods would be mixed. This raises concerns about the suitability of a public repository for sharing mass spectral data as reference data for the identification of chemical compounds detected by mass spectrometry. This may be the main reason for the continuing absence of a public repository of mass spectral data.

Although standardization of experimental methods of mass spectrometry is thought to be essential for sharing the mass spectral data of chemical compounds and standardized procedures to unify experimental protocols have been proposed, the metabolomics research community has not reached consensus on those proposals.[10,11] As research groups individually optimized their experimental methods based on their projects and the physicochemical properties of their target compounds, switching to other analytical methods would be almost impossible. Consequently, each group prepared its own reference mass spectral library by analyzing commercially available standard reagents. However, commercially available standard reagents, especially those of secondary metabolites produced by plants and microorganisms, are limited in number. Because this limited availability restricts the ratio of identified metabolites to those detected on LC-MS and -MS^2, it remains as low as 3–5% (48/1233) in plant[12] and 20–30% (175/626) in human tissues.[13]

Usually, metabolites are identified by comparing two data, retention index of chromatographic separation and mass spectrum, with authentic compounds analyzed under identical experimental conditions. New technologies such as single-cell mass spectrometry using matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry[14] and direct nano-ESI mass spectrometry[15] do not employ chromatographic separation but rather, they ionize all chemical compounds in a cell at once. Therefore, metabolite identification in new technologies depends solely on the reference library of the MS^n data.

In summary, although we must not expect the standardization of experimental protocols or platforms, this does not justify the absence of a public repository for mass spectral data.

Here, we report MassBank, the first public repository of mass spectral database of small chemical compounds (<3000 Da) for life sciences. Research groups contributing to the repository make their mass spectral data available to the public as supporting experimental data for other researchers. MassBank accepts mass spectral data analyzed on chemical compounds using optimized, up-to-date analytical methods. It is also the first internationally allied spectral database. As contributors deposit their mass spectral data not on a centralized repository, but on their own MassBank data servers, the contributed data and their quality are not mixed but independent from those of other contributors. Users of MassBank are provided with informatics tools to search the distributed data for identification of chemical compounds detected by mass spectrometry.

Experimental

Concepts underlying MassBank

We designed the software architecture and record format of MassBank based on three concepts. First, MassBank should be a public repository for sharing mass spectral data. Contributors should prepare their data in a common record format that defines the data field for the experimental methods, details the analytical parameters of the mass spectrometry and provides peak data. Second, data should be distributed on the Internet. Ideally, each contributor should have a local data server for publication of the formatted data. A contributor may have multiple databases to facilitate the separate management of data analyzed on different instruments, and (s)he could specify which data servers are and are not open to the public. Third, the query interface of MassBank functions as an access point to data servers distributed on the Internet.

Software architecture of MassBank servers

Despite its distributed design, from the user's point of view, MassBank should appear and function as a normal centralized database. Users should be able to access MassBank data without knowing where the data are or what data are involved and contributors should be able to update and manage their data independently.

Table 1. MassBank record

Tag	Description of record field
Summary section	
ACCESSION	Accession number
RECORD_TITLE	Short summary of the record, including the chemical name of the compound analyzed and the analytical method
DATE	Date of contribution
AUTHORS	Contributors and their affiliations
COPYRIGHT	Copyright notice
Chemical section	
CH$NAME	Chemical name of the compound analyzed
CH$COMPOUND_CLASS	Chemical class of the compound
CH$FORMULA	Chemical formula of the compound
CH$EXACT_MASS	Exact mass of the compound
CH$SMILES	SMILES code of the chemical structure of the compound
CH$IUPAC	InChI code of the chemical structure of the compound
Analytical section	
AC$INSTRUMENT	Mass spectrometer and name of manufacturer
AC$INSTRUMENT_TYPE	Type of ion analyzer
AC$ANALYTICAL_CONDITION/MODE	Ionization mode
Spectral section	
PK$NUM_PEAK	Total number of peaks
PK$PEAK	Peak data: *m/z*, intensity and relative intensity
Others	
MOLFILE_NAME	File name of the molfile that defines the chemical structure of the compound analyzed

Each data field is labeled by the tag specifying the data item. The 16 tags listed in the table are mandatory; they are shown on Record Editor.

To satisfy these requirements, we adopted a three-tier architecture for the MassBank system; it is comprised of database, application and presentation layers. The database layer stores the mass spectral data in text format in the relational MySQL database. The application layer is a search engine for the data stored in the database layer. The presentation layer is the user interface that specifies servers to be accessed. The application and presentation layers are implemented in Java on the Apache Tomcat web server.

Software distribution and maintenance

The MassBank system software is distributed free-of-charge under the GNU General Public License. The latest source codes are downloadable from SourceForge.net and they are provided for both Linux and Microsoft Windows operating systems (OS). MassBank Installer is a single archive file that includes precompiled object files and a script for the installation of required free software such as Apache, Tomcat and MySQL. As the MassBank Installer is not updated as often as the frequently updated MassBank system, we recommend that users install the MassBank system by means of the MassBank Installer first and then perform updates using the latest source codes from SourceForge.net.

An update service is provided to make maintenance of MassBank easy. The version of each component of the MassBank system is checked automatically using the http access to the MassBank.jp website. When an old component is found, the latest version is transferred and installed automatically.

MassBank record format

MassBank data must be prepared in the MassBank record format. Each record contains one mass spectrum attributable to one chemical compound with a specific chemical formula and each record consists of four sections: a summary, chemical, analytical and spectral section. Each data field carries a tag that specifies the data item (Table 1). For example, for the chemical, analytical and spectral sections the tags are CH$, AC$ and PK$, respectively.

The summary section contains the accession number that uniquely defines the record and summary information of the analytical and chemical sections, authors and copyright. The first three letters of the accession number specify the contributor.

The chemical section, CH$, defines the chemical information of the compound analyzed, including chemical names, the CAS number, compound category and IDs with links to available chemical compound databases such as KEGG,[16] PubChem,[17] KNApSAcK,[18] LipidBank,[19] and LipidMaps,[20] if available. The chemical structure is given in SMILES[21] and InChI code[22] and is defined separately by an MDL molfile.

The analytical section, AC$, describes the instrument types and analytical parameters used for mass spectrometry, including the instrument manufacturer, the catalog number of the mass spectrometer, the method of ionization, the type of ion analyzer, ionization voltage, matrix for MALDI ionization and the collision-induced dissociation (CID) conditions for MS^n measurement. For chemical compounds in biological samples that were separated and purified by LC, GC or capillary electrophoresis (CE) coupled to a mass spectrometer, the chromatographic column used, the chromatographic separation conditions and the retention index should be described in detail. These data are helpful for the identification of chemical compounds.

The spectral section, PK$, lists peak data with *m/z* and intensity and relative intensity values in integral or real numbers.

Evaluation of the precision of compound identification by spectral search

The query and target datasets (QSs, TSs) were prepared by extracting ESI-MS^2 data from MassBank data. The two datasets consisted of ESI-MS^2 data in which identical metabolites were analyzed under different analytical conditions. Using the QS spectrum as the query, a spectral search against TSs retrieved a list of similar spectra with corresponding similarity scores. If the metabolite of a similar spectrum was the same as the metabolite of the query spectrum, the search result was considered correct; if not, it was considered incorrect. Each search result was recorded with the similarity score. We repeated the spectral search for all QS spectra.

Considering the search results with a similarity score higher than the threshold, say s, to be true, we counted the number of true positives, TP(s), false negatives, FN(s) and false positives, FP(s), as follows.

TP(s) = Total number of correct results with a similarity score higher than the threshold value s,
FN(s) = Total number of correct results with a similarity score lower than the threshold value s,
FP(s) = Total number of incorrect results with a similarity score higher than the threshold value s.

We then calculated the precision, recall and *F*-value at threshold s as follows.

$$\text{Precision(s)} = \text{TP(s)}/[\text{TP(s)} + \text{FP(s)}] \quad (1)$$

$$\text{Recall(s)} = \text{TP(s)}/[\text{TP(s)} + \text{FN(s)}] \quad (2)$$

$$F\text{-value(s)} = \text{Harmonic means between Precision(s) and Recall(s)} \quad (3)$$

Results

Tools for contributors

Contributors to MassBank must prepare the mass spectral data in the MassBank record format and deposit the formatted data on their own MassBank data servers. Previously, data preparation involved tedious manual work. For example, for the analytical section, contributors had to manually detail the experimental methods and analytical parameters of mass spectrometry. Additionally, experience with MySQL and the Linux OS was essential for data management on their data servers. To reduce the workload and the experience requirement, we developed two tools: Record Editor and Administration Tool.

Generally, mass spectrometers output mass spectral data in the form of binary raw data readable only by the specific software provided by the instrument manufacturer. Binary raw data contain the peak data and the analytical parameters used to control the mass spectrometers. Previously, contributors had to manually extract the peak data and the analytical method, including parameters from the binary raw data with appropriate software. Then they manually prepared the data of the analytical and spectral sections in the MassBank record format.

The Mass++ program can directly import the binary raw data of major instrument companies and output the data in mzML and other data formats.[23,24] Mass++ has newly incorporated functionality that imports binary raw data and automatically outputs the spectral data and the analytical methods in the MassBank record format. The formatted data output from Mass++ is then combined with the molfile that defines the structure of the chemical compound in the Record Editor. This tool automatically calculates the chemical formula and the exact mass of the molecule, and generates SMILES and InChI codes to complete the chemical data section. After the accession number of the record, the authors and other necessary data are manually input in the summary section, and the Record Editor outputs a complete MassBank record as shown in Fig. 1.

Finally, using Administration Tool on a web browser, contributors can upload and manage their data on their MassBank data servers. Thus, contributors no longer need to have experience with either Linux or MySQL commands for data management.

Manuals are available from the manual page of the MassBank site (http://www.massbank.jp/en/manual.html) for contributors wanting to know more about Record Editor and Administration Tool.

Statistics of MassBank data

As of January 2010, 16 research groups, 12 in Japan, 3 in the United States and 1 in Germany, are contributing data to MassBank (Table 2). Mass spectral data, chemical compounds and analytical methods are summarized for each research group on the website (http://www.massbank.jp/en/published.html). These data are distributed on eight MassBank data servers, one of which is located in the Leibniz Institute of Plant Biochemistry (Halle, Germany). Eight small research groups currently without their own data servers contribute their data to the MassBank data servers in Japan or Germany. In January 2010, MassBank data included 10 294 mass spectra [9276 ESI-MS^n, 605 EI-MS, 137 fast atom bombardment (FAB)-MS] of 2337 chemical compounds, 3045 ESI-MS^2 data of 679 synthetic drugs and 11 545 EI-, 795 CI-, 38 FD- and 1 FI-MS data of 10 286 volatile natural and synthetic compounds. The MassBank data consist of data analyzed on 21 different instrument types.

MassBank data are composed of the mass spectra of primary metabolites, flavonoids, gibberellins, saponins, carotenoids, phospholipids and oligosaccharides. Most of these were analyzed on ESI-MS^2, and some on FAB-MS. In their analysis on ESI-MS^n, different CID energies were applied to obtain as many product ions as possible. This resulted in 9276 ESI-MS^n data of 1889 chemical compounds, an average of 4.9 ESI-MS^n data per chemical compound. EI-MS data are for bile acids and volatile chemical compounds such as terpenoids, alkyl alcohols, aldehydes and carboxylic acids. Since standard experimental conditions are available for EI-MS, each chemical compound has only one spectral datum.

In collaboration with LipidBank (http://www.lipidbank.jp/), the official database of the Japanese Conference on the Biochemistry of Lipids (JCBL), MassBank also collects the mass spectra of lipids from the literature. As of June 2008, MassBank is the official database of the Mass Spectral Society of Japan.

Users can access MassBank data from two access points, one in Japan[25] and the other in Germany.[26] Monthly access to MassBank data originating from Japan, USA, UK, Germany, Spain and other countries has reached 7800 hits on average, more than half originated from countries other than Japan.

Tools for users

Here we briefly introduce the tools developed for users to access MassBank data and their functions. Users wanting to know more details about the functions can consult a user manual available as a pdf file from the MassBank website (http://www.massbank.jp/en/manual.html).

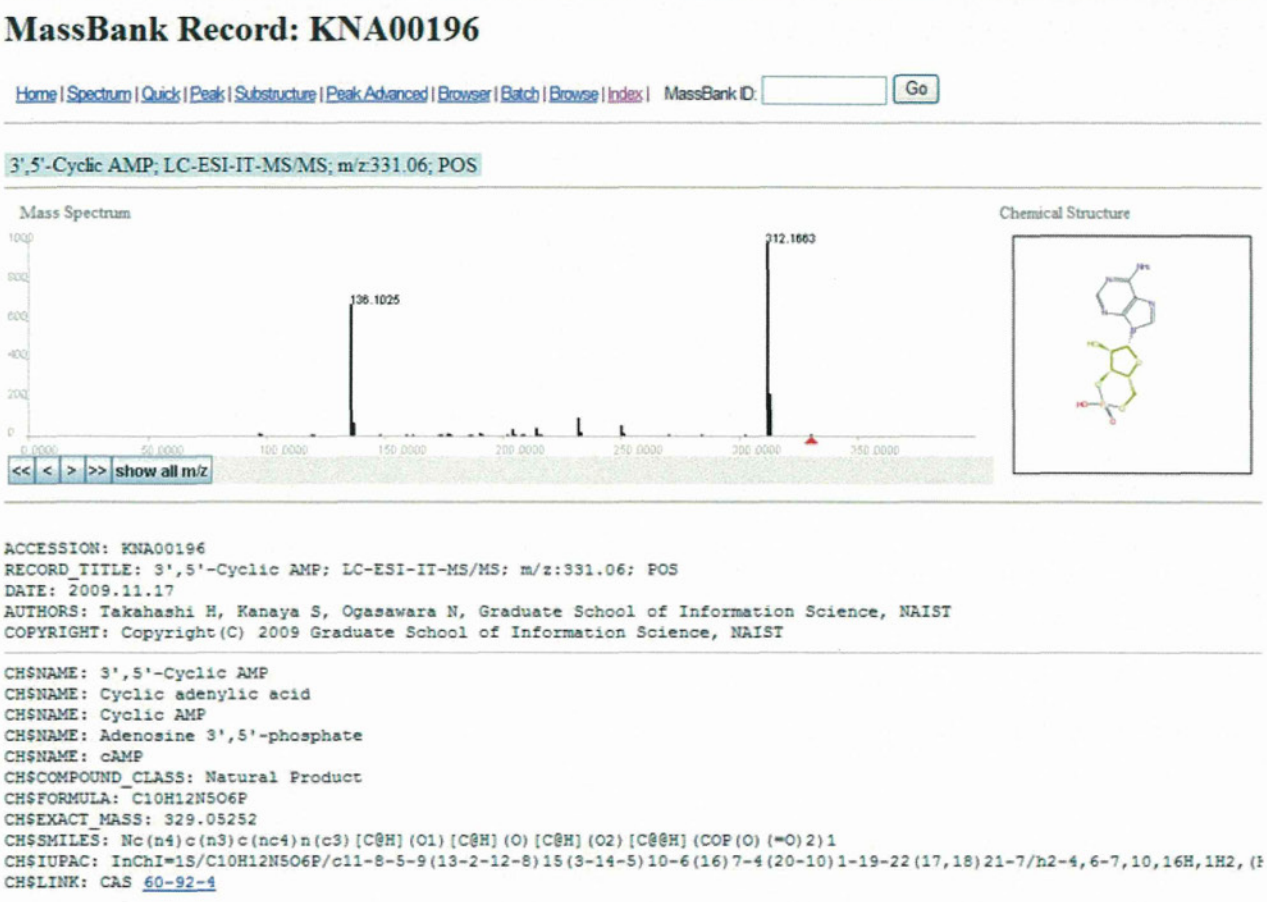


Figure 1. Example of a MassBank record.

To obtain suitable search results, users should specify search conditions using the Search Parameter Setting applet before their first search. The users should first specify the search tolerance, that is the experimental error allowance in the *m/z* value, the cutoff threshold for lower intensity peaks and the precursor ion by the *m/z* value. Then, the users select the instrument type identical with or similar to the type of the query mass spectrum and the ionization mode (Fig. 2(a)). Currently, the applet displays 21 instrument types.

Spectral Search

Spectral Search retrieves MS^n data identical with or similar to the query data. The search results are output in the order of the similarity score together with the number of identical product ions.

MassBank currently adopts the database search algorithm that calculates the similarity score based on a modified cosine correlation proposed by Stein and Scott.[27] The intensity of the *i*th peak is weighed by a factor, *W*i, as follows:

$$W_i = [\text{Intensity of peak}_i]^m [m/z \text{ of peak}_i]^n \quad (4)$$

Stein and Scott empirically determined the optimal exponents as $m = 0.6$ and $n = 3$ by analyzing *ca* 12 000 EI-MS data of 8000 organic compounds in the NIST Mass Spectral Library. Similar to their method, we optimized the exponents as $m = 0.5$ and $n = 2$ by analyzing 8785 ESI-MS^2 data of *ca* 700 authentic compounds of primary metabolites.[28] The difference between the present exponents and those determined by Stein and Scott is primarily attributable to the smaller number of peaks and the higher intensity of higher *m/z* peaks in the ESI-MS^2 data analyzed.

By displaying the search results peak-by-peak on the three-dimensional display, users can identify peaks in a database mass spectrum that are common to peaks in the query mass spectrum (Fig. 2(b)). MassBank provides a batch service for heavy users who submit many MS^n data as queries to the search service.

Quick Search and Substructure Search

MassBank features two tools to search for chemical compounds in its repository: Quick Search and Substructure Search. Quick Search retrieves chemical compounds by the chemical name, chemical formula and a list of the *m/z* and relative intensity values. The search results show the chemical compounds with their chemical names, spectral data and chemical structure (Fig. 3). Substructure Search retrieves chemical compounds containing a specified chemical substructure as a part of their chemical structure (Fig. 4). Users can select three different search options depending on how many π electrons in the query substructure are included in the target structures. The number of π electrons should be (1) the same, (2) higher in the target data or (3) ignored.

Peak Search and Peak Difference Search

Peak Search retrieves MS^n data containing the peaks specified by the *m/z* values within a specified error allowance. Peak Difference Search shows chemical compounds containing one or more peak pairs whose *m/z* values are different from each other by the specified *m/z* values.

Table 2. Statistics of MassBank data as of January 2010

Research group	Group ID	Analytical method	Num of spectra	Num of compounds
Institute for Advanced Biosciences, Keio University	KO	ESI-QqTOF-MS/MS	914[a]	695
		ESI-QqQ-MS/MS	4 275	
		ESI-IT-$(MS)^n$	515	
PSC, RIKEN	PR	GC-EI-TOF-MS	241	767
		LC-ESI-TOF-MS	85	
		LC-ESI-QqQ-MS/MS	87	
		CE-ESI-TOF-MS	20	
		LC-ESI-QTOF-MS/MS	1 290	
Waters	WA	LC-ESI-Q-MS	2 721	577
		ESI-QqQ-MS/MS	273	
Akimoto, Graduate School of Pharmaceutical Sciences, Kyoto and Maoka, Research Institute for Production Development	CA	FAB-CID-EBEB-MS/MS	106	106
Taguchi, Graduate School of Medicine, The University of Tokyo	UT	ESI-QqIT-MS/MS	378	42
Kazusa DNA Research Institute	KZ	GC-EI-TOF-MS	273	163
Iida, College of Humanities and Sciences, Nihon University	NU	EI-MS	75	74
Tanaka, Institute of Natural Medicine, University of Toyama	TY	LC-ESI-IT-TOF-MS	91	69
Kimura, Faculty of Agriculture, Tottori University	TT	EI-MS	11	11
		FAB-MS	5	
Funatsu, Graduate School of Engineering, The University of Tokyo	JP	EI-MS	11 545	10 286
		CI-MS	795	
		FD-MS	38	
		FI-MS	1	
Leibniz Institute of Plant Biochemistry	PB	ESI-QqTOF-MS/MS	297	90
		ESI-QqQ-MS/MS	63	
Matsuura, Fukuyama University	FU	LC-ESI-QqQ-MS/MS	285	71
Metabolon, Inc.	MT	ESI-IT-MS/MS	149	149
Morii, University of Occupational and Environmental Health	UO	FAB-MS	26	25
		EI-MS	5	
		FD-MS	3	
		CI-MS	1	
Kanaya, Graduate School of Information Science, Nara Institute of Science and Technology	KNA	LC-ESI-IT-MS/MS	619	75
		LC-ESI-FT-MS	208	
Grant, University of Connecticut	CO	ESI-QqTOF-MS	510	102

[a] Number of merged spectra.

Peak Search Advanced

Peak Search Advanced is similar to Peak Search and Peak Difference Search in function, but it is different in that it specifies the peaks with the molecular formulae of the ions. Peaks in the merged data (see the next section for details) are annotated by the chemical formula within an error range of 50 ppm (the threshold is adjustable). Currently, there are 817 positive and 797 negative ESI-QqTOF-MS^2 merged data available as the target for Peak Search Advanced.

Merged mass spectra as artificial reference mass spectra for metabolite identification

One of the most important applications of MassBank data in the life sciences is metabolite identification. Generally, ESI-MS^2 data of chemical compounds are useful as reference data for metabolite identification when the analytical conditions of the query ESI-MS^2 data are the same as or very similar to those of the reference mass spectra. When the query and the reference chemical compounds are the same, the spectral search retrieves the reference mass spectrum with higher similarity scores. In other cases, the query and reference mass spectra are less similar or different even when the two chemical compounds are the same. As most MassBank users may encounter the latter situation, MassBank provides an artificial reference, that is the 'merged' mass spectrum.

As the reproducibility of the ESI-MS^2 data is reportedly low,[29,30] we evaluated the degree of reproducibility of MassBank ESI-MS^2 data for use as reference data in the metabolite identification. We took two datasets of common metabolites extracted from MassBank: datasets [QqQ] and [QqTOF] consisting of 4205 ESI-QqQ-MS^2 and 4431 ESI-QqTOF-MS^2 data of 856 common chemical compounds, respectively. Each chemical compound in each dataset has four or five spectral data. In the first experiment,

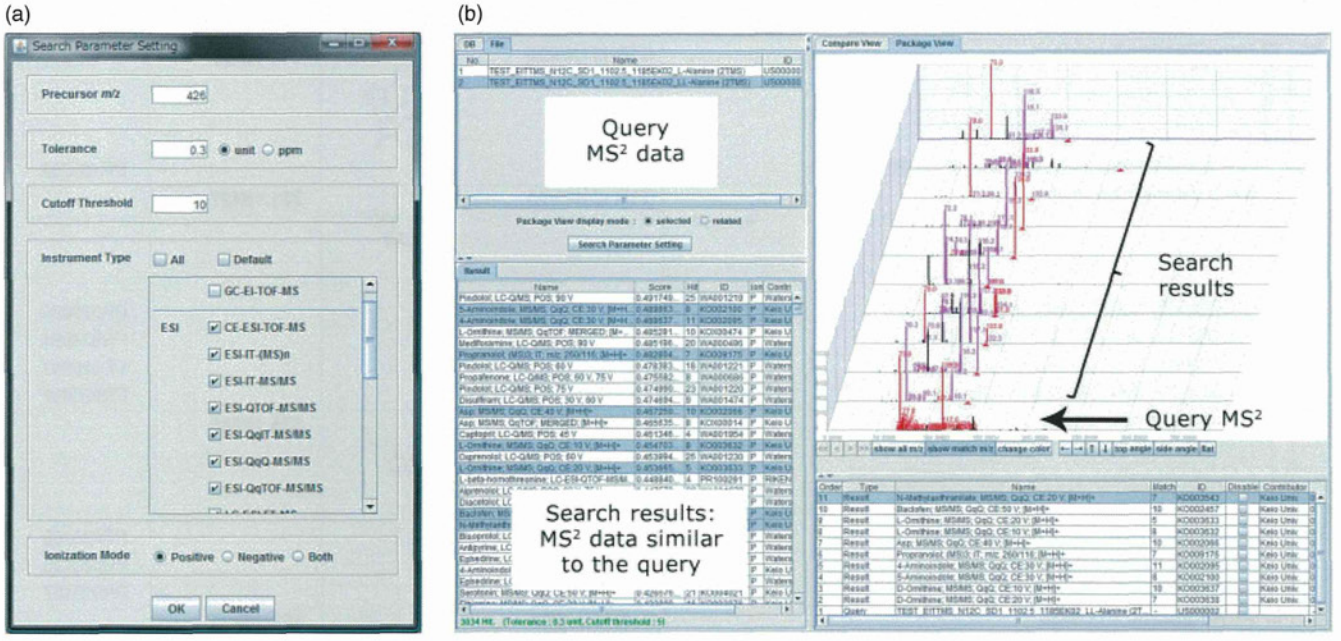


Figure 2. Search Parameter Setting and Spectral Search. (a) Search parameters are selected and input on the applet. The 'Precursor ion' is specified by the *m/z* value. 'Tolerance' is the error allowance of *m/z* values. When a peak in the query data and the corresponding peak in the target data have different *m/z* values but are within the tolerance, the two peaks are treated as identical. 'Cutoff threshold' is used to distinguish real peaks from noise peaks. (b) The left upper and lower panels show the QS and the search results in the order of the similarity score, respectively. When some of the search results are selected in the left lower panel, the three-dimensional display in the right upper panel shows the spectral search results in peak-by-peak mode.

the query dataset (QS) was [QqQ] and the TS was [QqTOF]. In the second experiment, QS and TS were [QqTOF] and [QqQ], respectively. We performed two spectral searches and evaluated precision (see Experimental section, Eqn (1)), recall (Eqn (2)) and the *F*-value (Eqn (3)) at various threshold similarity scores for each QS and TS pair. When the threshold of the similarity score was 0.6, the precision, recall and the *F*-value for TS = [QqQ] and [QqTOF] were [0.222, 0.327, 0.264] and [0.276, 0.292, 0.284], respectively. Thus, in their original form, ESI-MS^2 data in MassBank are not likely to serve as reference data.

ESI-MS^2 data using CID reflect the employed collision energy (Fig. 5(a)); smaller product ion nonlinearly increase with the collision energy. This is one of the major reasons for the low reproducibility of ESI-MS^2 data analyzed under different analytical conditions. Therefore we expect that merged mass spectra, that is superposition of spectra in different collision energies, would better serve as the reference mass spectra for metabolite identification.

In fact, metabolomics groups at the Institute for Advanced Biosciences, Keio University, Tokyo, Japan ('Keio group') and the RIKEN Plant Science Center, Yokohama, Japan ('RIKEN group') measured the ESI-MS^2 data of chemical compounds at five different CID collision energies in both positive and negative modes. The Keio group assessed 4570 ESI-QTQF-MS^2 data of 695 chemical compounds under five different collision energies at 10–50 V. For each chemical compound, the ESI-QTQF-MS^2 data were overlaid and merged into a single artificially merged MS^2 spectrum (Fig. 5(b)). Each of the chemical compounds has one merged mass spectrum. The Keio group contributed 914 merged ESI-QTQF-MS^2 data of 695 chemical compounds to MassBank. The RIKEN group measured 535 chemical compounds on LC-ESI-QTOF-MS^2 under the ramp mode, which we regard as merged mass spectra, in the range of 5–60 V collision energies in both positive and negative modes, contributing to a total of 1290 ESI-MS^2 data. Merged mass spectral data have the character 'X' in the third position of the record number, e.g. KOX000031. These merged ESI-QTOF-MS^2 data contain most of the product ions observed under the commonly adopted CID conditions for measuring ESI-MS.[2] Therefore, for each chemical compound, the merged data yield a representative fragmentation pattern.

Evaluation of compound identification using merged ESI-MS^2 data as reference data

We evaluated the quality of merged ESI-MS^2 data as reference data vis-à-vis the original ESI-MS^2 data. The TSs [Merged QqQ] and [Merged QqTOF] were prepared by merging [QqQ] and [QqTOF] for each chemical compound. This yielded 856 merged data for each dataset. In the first experiment, QS was [QqQ] and TS was [Merged QqTOF], and in the second, QS and TS were [QqTOF] and [Merged QqQ], respectively. We performed two spectral searches and evaluated precision, recall and the *F*-value at various threshold similarity scores for each QS and TS pair. When the threshold of the similarity score was 0.6, precision, recall and *F*-value observed for TS = [Merged QqQ] and [Merged QqTOF] were [0.454, 0.307, 0.366] and [0.490, 0.299, 0.371], respectively. Therefore, merging the ESI-QqQ and QqTOF-MS^2 data improved the precision of the spectral searches by 23% and 21%, respectively, at similarity scores higher than 0.6. Merging the data did not significantly affect recall. The merged data improved metabolite identification using ESI-QIT-MS data as queries (data not shown). Therefore, a spectral search with weighting parameters optimized against the merged mass spectra yields satisfactory results for metabolite identification.

We recommend that contributors of ESI-MS^2 data deposit multiple data for each chemical compound analyzed under at least a few different levels of collision energy in both positive and negative mode.

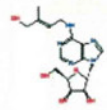
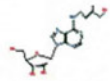
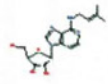
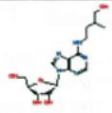
Name	Formula / Structure	ExactMass	ID
trans-Zeatin-riboside 1 spectrum	C15H21N5O5 	351.15427	
LC-ESI-QTOF-MS/MS; CE:Ramp 5-60 V; [M+H]+			PR100209
trans-Zeatin riboside 4 spectra	C15H21N5O5 	351.15427	
LC-MS/MS; QqQ; CE:40.0 eV; [M+H]+			PR020095
LC-MS/MS; QqQ; CE:30.0 eV; [M+H]+			PR020094
LC-MS/MS; QqQ; CE:20.0 eV; [M+H]+			PR020093
LC-MS/MS; QqQ; CE:10.0 eV; [M+H]+			PR020092
isopentenyladenosine 3 spectra	C15H21N5O4 	335.15935	
LC-MS/MS; QqQ; CE:30.0 eV; [M+H]+			PR020109
LC-MS/MS; QqQ; CE:20.0 eV; [M+H]+			PR020108
LC-MS/MS; QqQ; CE:10.0 eV; [M+H]+			PR020107
dihydrozeatin riboside 3 spectra	C15H23N5O5 	353.16992	
LC-MS/MS; QqQ; CE:30.0 eV; [M+H]+			PR020104
LC-MS/MS; QqQ; CE:20.0 eV; [M+H]+			PR020103
LC-MS/MS; QqQ; CE:10.0 eV; [M+H]+			PR020102

Figure 3. Quick Search. When, for example, the search involves chemical compounds containing 'adenine' in the name, Quick Search displays the chemical compounds matching the search together with the spectral data and chemical structure.

API services

The MassBank Application Programming Interface (API), the Simple Object Access Protocol (SOAP) interface to MassBank, allows users to write their own programs for accessing, customizing and utilizing MassBank. Currently available methods, downloadable from http://www.massbank.jp/en/download.html and described by a schema in Web Service Definition Language (WSDL) (http://www.massbank.jp/api/services/MassBankAPI?wsdl), are Spectral Search, Peak Search and Peak Difference Search.

We show an example using MassBank API. As described above, mass spectrometers output spectral data as binary raw data. Because binary raw data are not accepted as a query for a spectral search in MassBank, they must first be converted into text data format. Conducting a spectral search query for several hundred binary raw data outputs with a single run of LC-MS^n was a time-consuming task in metabolomics studies. The Mass++ program frees users from this burden with a new function that imports binary raw data for submission as a spectral search query using MassBank API and shows the search results in its own display mode. In the near future, MassBank will provide the WSDL batch service method for spectral searches.

Program source codes and tool manuals

MassBank is currently available in Linux and Microsoft Windows versions. Typically, the Windows version is released more than 6 months after the Linux version. The source codes of the MassBank system are freely available from SourceForge[31] with the GNU General Public License. Manuals for using the search tools, preparing the data in the MassBank record format, installing the MassBank system and for managing data on MassBank servers are available from the MassBank Manual download site.[32]

Discussion

Merged mass spectra for the identification of chemical compounds

Public mass spectral databases accept mass spectral data analyzed by nonstandardized analytical methods. Among different analytical methods, ESI-MS^n data are of low reproducibility; therefore, these data were not thought to be useful as reference data. However, Volná *et al.*[30] found that the fragmentation patterns are almost identical for all tandem mass analyzers and that only the ratios of the product ions differ somewhat. They recommend analyzing ESI-MS^n at three different CID collision energy levels. Our present analysis of MassBank data supports their findings. In fact, most contributors of ESI-MS^n data to MassBank analyzed each chemical compound under five collision energy levels ranging from 5 to 50 V to observe all possible product ions. Additionally, MassBank provides a merged mass spectrum for each compound. Although merging ESI-MS^n data statistically improved the precision of metabolite identification without decreasing recall, we encountered two problems with the merged data. First, the total number of product ions in the merged data tended to be much larger than the number of product ions in the original ESI-MS^n data. For example, merging five data increased the total number of product ions by 3.82 times (an average of 870 merged data). This resulted in an increase in the number of false-positive hits and a consequent decrease in precision. Second, the base

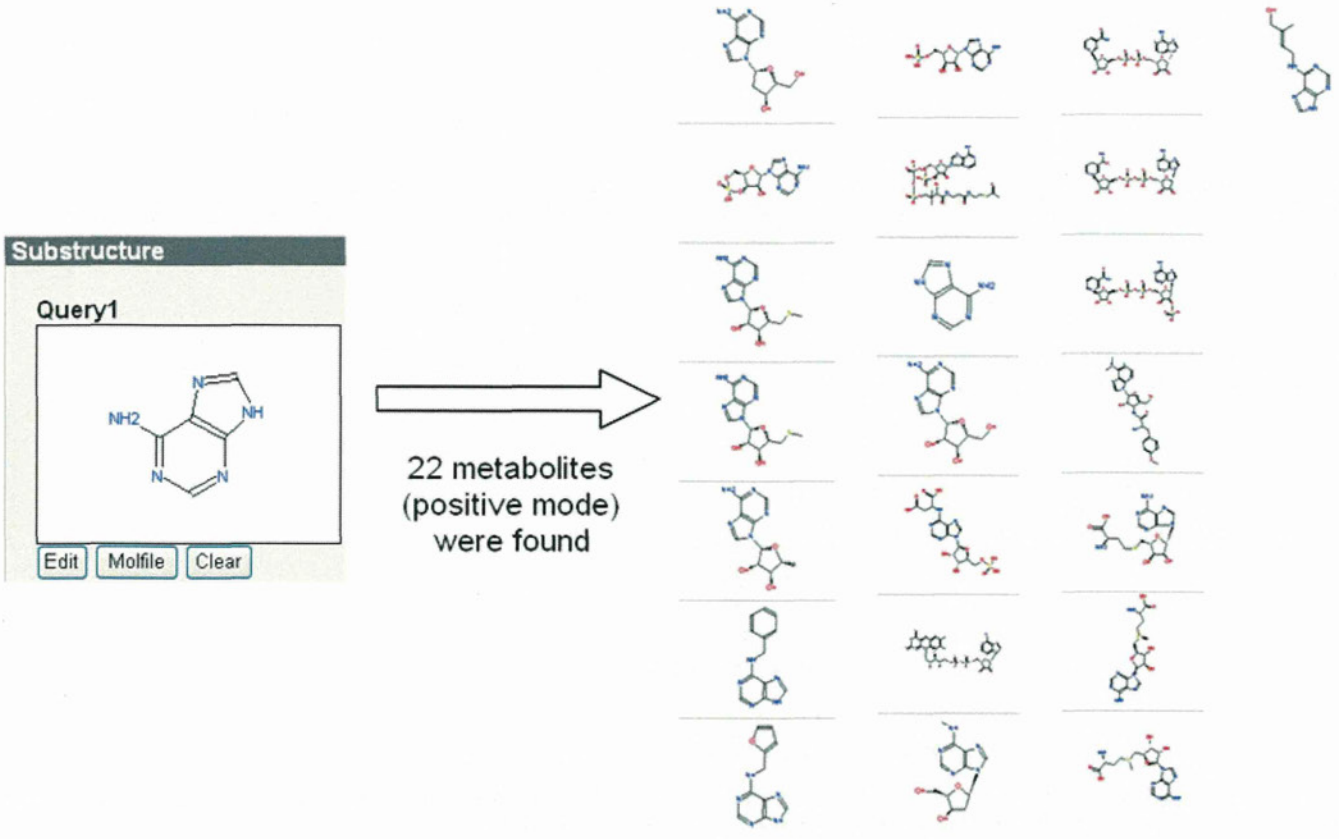


Figure 4. Substructure Search. When a substructure is submitted as a query, all chemical structures containing the query substructure are listed.

peak in the merged data was different from the base peak in the original data. The development of a better merging method and a new database-searching algorithm will solve these problems and improve metabolite identification in MassBank.

Cost of publication of a distributed database

In MassBank, contributing research groups openly avail their data to the public from their own data servers. From this aspect, MassBank is similar to the currently available mass spectral databases discussed in the Introduction (GMD@CSB.DB, METLIN, GMDB, HMDB, NIST/EPA/NIH Mass Spectral Library, SDBS). However, MassBank is different because it accepts data contributions from researchers and groups; the repository contains data analyzed with a wide range of mass spectrometry methods. Via the Search Parameter Setting interface, MassBank allows users to select datasets obtained with different analytical methods as the search target.

In other databases, only the owning research groups or laboratories contribute to their databases and the data in each database are prepared in different record formats. Consequently, the (owning) users of a database cannot access other (nonowned) databases in parallel. In MassBank, contributors must prepare their data in the specified record format. This includes not only the peak data but also the analytical method and conditions, and the chemical structure information on the analyzed chemical compounds. In addition, contributors must manage their data on their own local data servers. As the preparation of formatted data and data management on owned servers was time consuming, at the request of contributors we made efforts to reduce their workload. Our efforts resulted in an increase in the data deposited in MassBank in 2009.

The cost incurred by contributors in the preparation and management of their data in the MassBank-distributed database system is proportional to the amount of data deposited. Contributors of larger quantities of data need high-performance computers and large storage capacity. This is one of the rationales behind a distributed database system. In grant applications, contributors should include costs involved in the publication of experimental data as a necessary expense for the sharing of their data as a research product. Funding organizations should judge the performance of researchers not only based on publications but also on products made available to the wider research community.[1]

The freely available source code is also useful for an independent database project outside of the MassBank consortium. An example is MS/MS spectral tag (MS2T) viewer[12,33] where the data are prepared in the MassBank record format and whose database server is the MassBank clone. The users cannot access the viewer from the common MassBank interface, but only from its original website (http://prime.psc.riken.jp/lcms/ms2tview/ms2tview.html).

Retaining the quality of mass spectral data in MassBank

Some users of MassBank are concerned with the quality of MassBank data with respect to the technical quality of the mass spectrometry and the chemical purity and identification levels of the samples. At present, we cannot offer a practical method for evaluating the technical quality of contributed data. However, before data submission, contributors can easily look for experimental mistakes on Record Editor. Thus, mistakes such