

Introduction

Software Dr DMASS has been developed to effectively analyze mass spectral data based on multivariate analysis. Figure 1 shows a flow diagram of Data processing consisting of three stages, (i) Peak Correction, (ii) Multivariate Data Processing, and (iii) Multivariate Analysis. In Peak Correction process, we correct experimental m/z values based on the relation between experimental and desired values in internal mass calibrants (**IMCs**). A multivariate data is constructed by a data set of multiple samples. In Multivariate Data Preprocessing, we can assess reproducibility of samples with iterative measurement, and select useful peaks for separating groups of samples and so on. After preprocessing, we can visualize the multivariate data by using multivariate analysis method such as principal component analysis (PCA) and Batch-learning self-organizing map (BL-SOM).

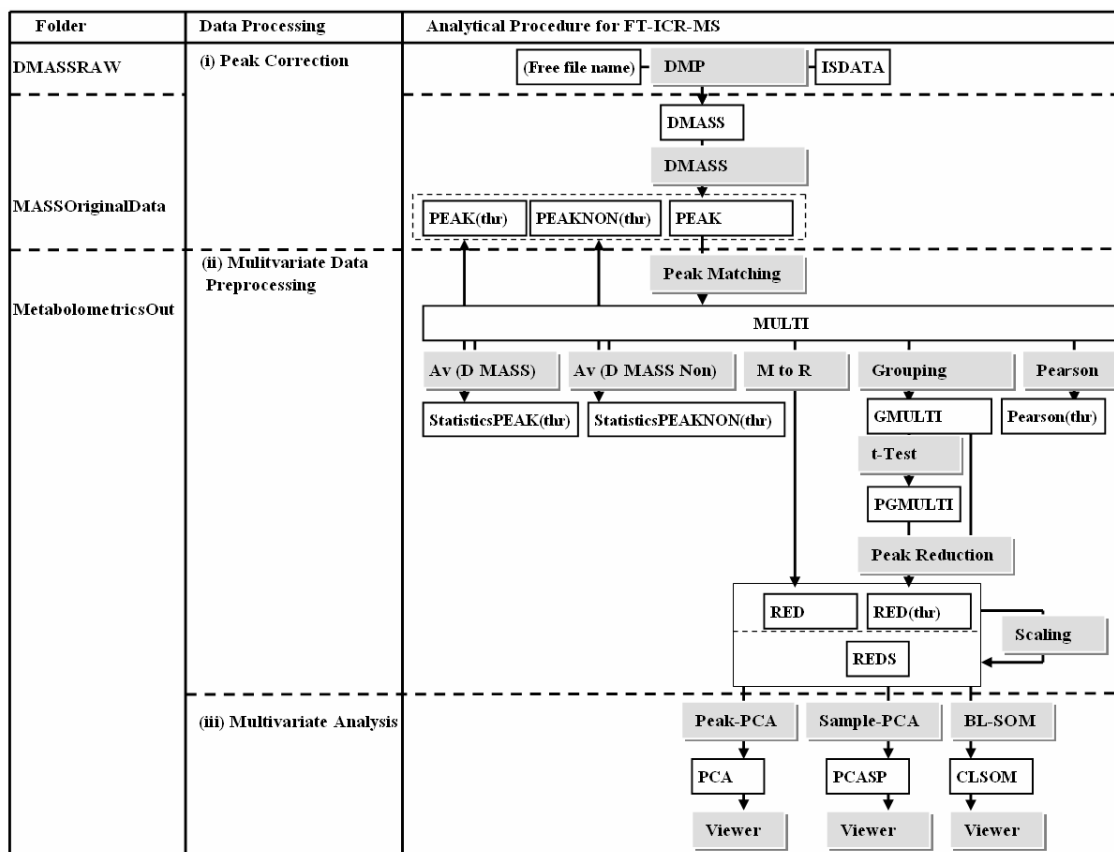
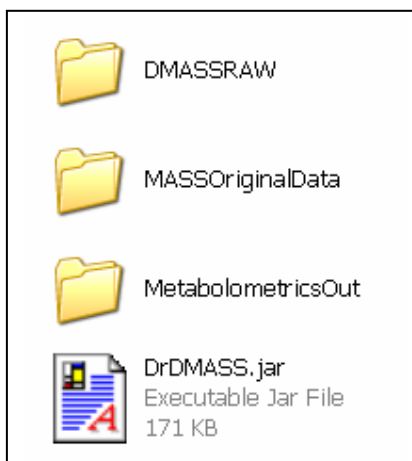


Figure 1 Flow Diagram of Data Processing in Dr DMASS (Silver boxes correspond to individual processes, and white boxes correspond to prefix in input/output file names.)

1. Execution of Dr DMASS

Java j2sdk-1.4.2 is required to be installed in the user's computer. First, the compressed file, DrDMASS.zip is to be downloaded from <http://kanaya.aist-nara.ac.jp/DrDMASS/>. Under the 'DrDMASS' folder, there are three folders 'DNASSRAW', 'MASSOriginalData', and 'MetabolometricOut', and an executable file 'DrDMASS.jar'.



1.1 Starting Data files

Put Digital mass spectral data and an internal mass calibrant data to 'DMASSRAW' folder. The calibrant data file name should start with 'ISDATA'. These file formats are as follows.

Digital mass spectral data

Digital mass spectral data from an IonSpec Explorer FT-ICR (IonSpec Inc., Lake Forest, CA) equipped with a 8 tesla actively shielded super conducting magnet is a text file separated by tabs. The first to fifth columns correspond m/z, Frequency, Amplitude, Relative abundant and Resolution.

m/z	Frequency	Amplitude	Rel.Abund.	Resolution
72.9895	1475427.93	0.0832	1.81	144100
73.6554	1462089.917	0.045	0.98	189100
.....				
.....				

ISDATA

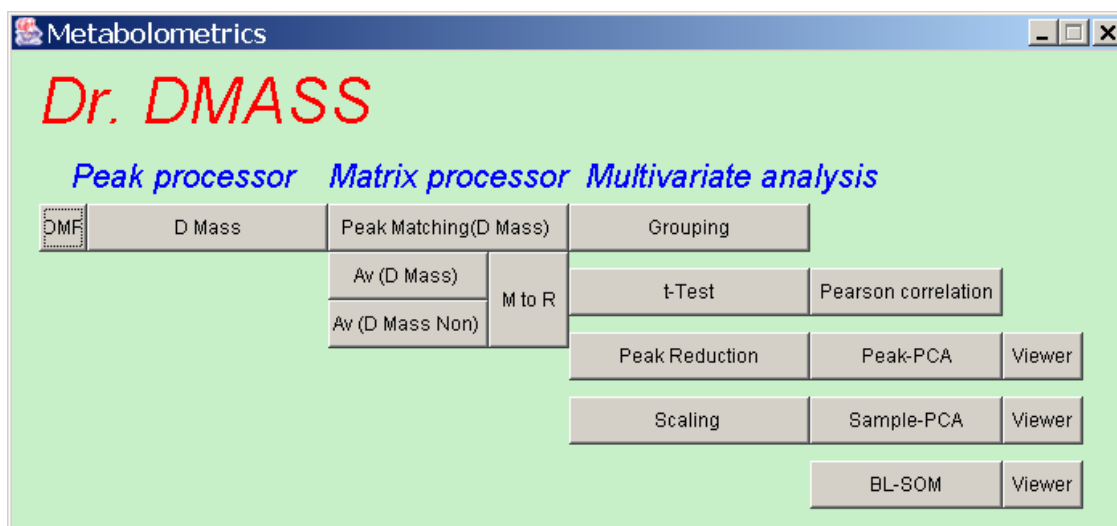
ISDATA consists of m/z values for internal mass calibrants (**IMCs**).

218.96212
348.10235
613.38820
829.32078

1.2 Execution of DrDMASS

User can start by clicking the file DrDMASS.jar. The main window is shown in Panel 1. The button names correspond to those in Figure 1. DrDMASS consists of 14 Data processing modules corresponding to the button names. (The summary of the processes is as follows.) In the present system, the processed single mass spectral data is put into 'MASSOriginalDATA', and the multivariate data is put into 'MetabolomicsOut' folder. Prefix for each input/output file is described in Figure 1. Each process is explained in the next section.

- | | |
|--------------------------------------|---|
| 1. DMP | Selection of m/z values for IMCs from Digital mass spectral data |
| 2. DMASS | Correction of m/z values for all peaks by those for IMCs. |
| 3. Peak Matching (D MASS) | Matrix construction for multiple samples
m/z-value and intensity for each sample is arranged to the matrix |
| 4. Av (D MASS) | Calculation of average intensity for all samples |
| 5. Av (D Mass Non) | Calculation of average intensity for samples with non-zero intensity. |
| 6. M to R | Construction of multivariate data consisting of m/z values and the intensity of multiple samples |
| 7. Group | Definition of categories for individual samples |
| 8. t-Test | Estimation of p-values by t-statistics for the difference between the average intensities for pairs of groups |
| 9. Peak Reduction | Selection of peaks with the group differences by p-values |
| 10. Scaling | Scaling data |
| 11. Pearson correlation | Pearson correlations of the intensities for pairs of m/z larger than the threshold set by the user are list up. |
| 12. Peak-PCA and its Viewer | Principal component analysis for peaks and visualization of its results |
| 13. Sample-PCA and its Viewer | Principal component analysis for samples and visualization of its results |
| 14. BLSOM and its Viewer | Batch-learning SOM for peaks and visualization of its results |



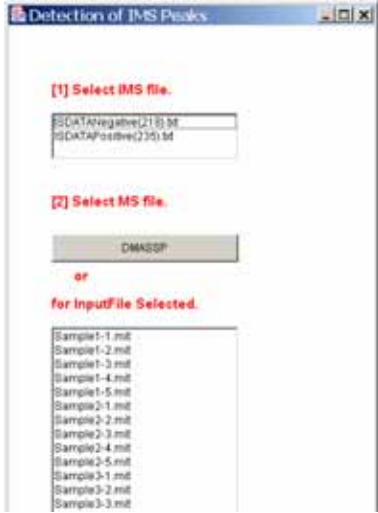
Panel 1 The main window

2 Explanation of individual processes

(i) Peak Correction

2.1. DMP

DMP process is a selection process of m/z values for IMCs from digital mass spectral data. The nearest m/z value in the digital mass spectral to those for IMCs is selected.

Input file	(i) Digital spectral data (its naming is free), and (ii) ISDATA
Output file	DMASS
Execution	<p>[1] Click DMP button, so the following panel is displayed. [2] Select an suitable IMC file consisting of m/z values in internal mass calibrants. [3] Click DMASSP button if the selection process of m/z values for IMCs is carried out for all MS files, or click an inputfile name (for example Sample1-1.mit) if the selection process is carried out for a targeted input file. In the demonstration data, we select ISDATANegative(218).txt and click 'DMASSP' button.</p> 
Output file format (DMASS)	<pre>>Sample1-1.mit Standard 218.9664 218.96212 348.1119 348.10235 613.4106 613.3882 829.3563 829.32078 // AllData 72.9895 0.0832 73.6554 0.045 976.4644 0.0313 //</pre> <p>1st line represents Inputfile name. From 2nd line to '/' (7th line): m/z values for IMCs are listed, that is, experimental and theoretical values for individual IMCs correspond to the first and second columns. From AllData to '/' (final line), m/z and its intensity are arranged.</p>

2. 2 DMASS

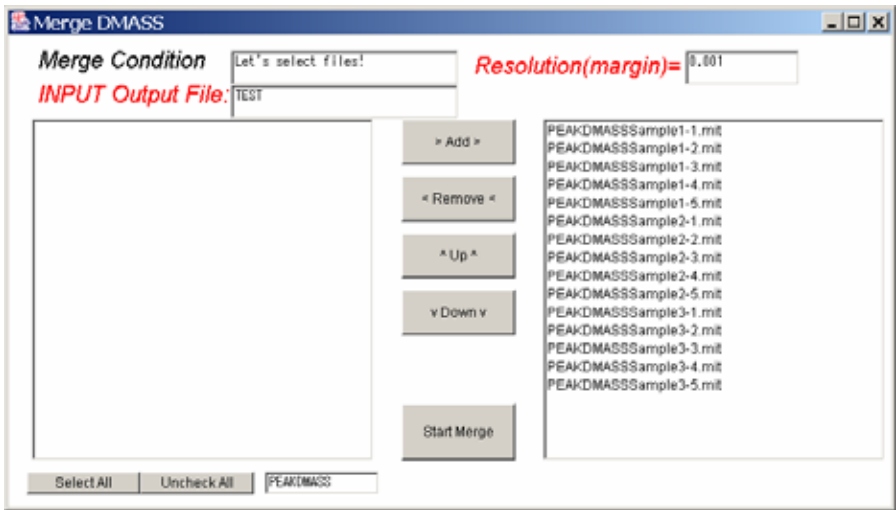
All m/z-values are corrected by linear relationship between theoretical and experimental values in the interval defined by the nearest m/z-values for IMCs.

Input file	DMASS
Output file	PEAK
Execution	[1] Select Samples used by clicking any number of filenames or 'Select all' button . [2] Click 'Start Correct' button . output files started with 'PEAK' are obtained.
Output file format (PEAK)	<pre> >Sample1-1.mit Standard 218.96212 218.96212 348.10235 348.10235 613.3882 613.3882 829.32078 829.32078 // AllData 72.99117683367211 0.0832 73.65704966049918 0.045 976.4199422981981 0.0313 // </pre> <p>1st line represents Inputfile name. From 2nd line to '/' (7th line), m/z values shown for IMCs are listed, that is, the theoretical values for individual IMCs correspond to both columns. From AllData to '/' (final line), corrected m/z and its intensity are arranged.</p>

(ii) Multivariate Data Preprocessing

2.3 Peak Matching (D MASS)

According to m/z values, peaks for multiple samples are arranged to matrix.

Input file	PEAK
Output file	MULTI
Execution	 <p>[1] Input Output file name. [2] Input Resolution (margin). This parameter determine the region of m/z as identical positions. [3] Select Samples used by clicking any number of filenames or 'Select all' button. [4] Click 'Add' button, then the selected filenames are moved from left to right side. [5] Prepare the order of filenames by selecting a filename and using 'up' and 'down' buttons. The order of filenames corresponds to that the order from left to right in the constructed matrix. [4] Click 'Start Correct' button. output files started with 'MULT' are obtained.</p>
Output file format (MULTI)	<pre>Resolution=0.0010 >Sample1-1.mit >Sample1-2.mit ... >Sample1-1.mit standard[0] 218.96212 218.96212 218.96212 218.96212 standard[3] 829.32078 829.32078 829.32078 829.32078 ... 63.67552147 0.0 0.0 0.0 0.0 ... 72.99098211 72.991176 0.0832 72.98792 0.0912 976.3983668 976.41994 0.0313 976.3946 0.0353 ... //</pre> <p>1st line represents Resolution. Lines started with '>' are Sample names analyzed. standard[0] to standard [3] represent corrected m/z for IMCs. After IMC line, pairs of m/z and its intensities are arranged according to the order of Sample names.</p>

Statistics of Av (D MASS)

In M iterative measurements, the intensities for IMCs and m/z are represented by data matrices \mathbf{Y} and \mathbf{X} , respectively. Here, the number of IMCs are denoted by S, and the number of peaks denoted by N.

$$\mathbf{Y} = \begin{pmatrix} y_{11} & \dots & y_{1j} & \dots & y_{1M} \\ \dots & \dots & \dots & \dots & \dots \\ y_{i1} & \dots & y_{ij} & \dots & y_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ y_{s1} & \dots & y_{sj} & \dots & y_{sM} \end{pmatrix} \quad (2.4.1)$$

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ x_{N1} & \dots & x_{Nj} & \dots & x_{NM} \end{pmatrix} \quad (2.4.2)$$

Line that starts with 'AvRef' in StatisticsPEAK(thr) corresponds to the average of IMCs for jth measurement represented by Eq.(2.4.3).

$$\bar{y}_j = \frac{\sum_{i=1}^S y_{ij}}{S} \quad (2.4.3)$$

The values under the columns Av, SD, Av/SD correspond to \bar{y} represented by Eq. (2.4.4), $SD(\bar{y})$ represented by Eq. (2.4.5), and $\bar{y}/SD(\bar{y})$, respectively.

$$\bar{y} = \frac{\sum_{j=1}^M \bar{y}_j}{M} \quad (2.4.4)$$

$$SD(\bar{y}) = \frac{\sqrt{\sum_{j=1}^M (\bar{y}_j - \bar{y})^2}}{M - 1} \quad (2.4.5)$$

Line that starts with 'TOTAL' corresponds to the statistical parameters for the intensities in jth measurement in a set of m/z whose intensities for all measurements are not zero is denoted by \bar{x}'_j .

Concretely, $(total(x'_j))$ corresponds to total intensities for jth.

Line 'CRTOTAL' corresponds to statistical parameters for corrected intensities by the average of IMCs, that is,

$$cav(x'_j) = \frac{av(\bar{x}'_j)}{\bar{y}_j} \quad (2.4.6)$$

Here,
$$av(x'_j) = \frac{\sum_{j=1}^M total(x'_j)}{N'}$$
.

In 'Threshold' line, threshold set by the user and the number of peaks satisfied by this condition are represented. From the following line to the last line, m/z value, corrected intensity for each sample are calculated as represented in Eq. (2.4.7).

$$c(x'_{ij}) = \frac{x'_{ij}}{y_j} \quad (2.4.7)$$


The column 'nonzero' corresponds to the number of intensities larger than zero, correctAv corresponds to the average of $c(x'_{ij})$ represented by Eq. (2.4.8)

$$av(x'_i) = \frac{\sum_j c(x'_{ij})}{M'_i} \quad (2.4.8)$$

Here M'_i represents the number of measurements larger than zero for i th m/z. The m/z and its correctedAv is arranged in the other output file (PEAK(thr)).

2.6 M to R

In 'M to R', a multivariate data matrix consisting of average m/z values and the intensities for multiple measurements is constructed.




Input file	MULTI
Output file	RED (data format for multivariate analyses in Dr DMASS system)
Execution	<p>Click filename, then format in MULT is exchanged to that in RED file.</p> 
Output file format (RED)	<p>1st line corresponds to inputfile name and 2nd and 3rd lines correspond to group index and merged filenames, respectively. 4th to the last lines corresponds to m/z and intensities for individual measurements.</p> <pre> >MULTITEST.txt >no. 1 2 3 ... >filename >Sample1-1.mit >Sample1-2.mit >Sample1-3.mit ... 72.99098211 0.0832 0.0912 0.0963 ... 73.65656016 0.0450 0.0535 0.0571 </pre>

Feature (m/z) Selection

Peak selection of the statistical significant differences in intensities between groups is carried out by the sequential processes entitled Group (Section 2.7), t-Test (Section 2.8), and Peak Reduction (Section 2.9).

2.7 Group

Definition of categories for individual samples

Input file	MULTI
Output file	GMULTI
Execution	<p>[1] Click Grouping button, then Group DMASS panel is displayed.</p> <p>[2] Select filename, then 'Select files to group' panel is displayed.</p>   <p>[3] Select files belonging to the same group, and move these files to right side by clicking 'add' button.</p>  <p>[4] Click 'Decide' button. So selected file is removed and their group is assigned to output file started with 'G'.</p>



[5] Continue the manipulation of [3] and [4] till groups for all files are assigned, that is, all files are removed.

[6] Click 'Start Grouping' button, then grouping process is started.




Output file format

Lines started with '>group' correspond to the numbers of individual groups. The line with '>no' corresponds to group ID, the line with 'filename' corresponds to the labels for individual measurements, and the m/z-values and their intensities are arranged.

```
>group1    5
>group2    5
>group3    5
>no.       1      ...      1      2      ...      2      ...
>filename  Sample1-1.mit ...      Sample1-5.mit  Sample2-1.mit ...
72.99098211 0.0832 ...      0.0836  0.0635 ...      0.0704 ...
73.65656016 0.0450 ...      0.0496  0.0477 ...      0.0419 ...
77.07104731 0      ...      0      0      ...      0      ...
....
....
```


2.8 t-Test

Estimation of p-values by t-statistics for the difference between the average intensities for pairs of groups

Input file	GMULTI
Output file	PGMULTI
Execution	<p>Click filename, then p-values by t-statistics for the difference between the average intensities for pairs of groups are calculated for individual m/z.</p> 
Output file format	<p>Lines started with '>group' correspond to measurements belonging to the same groups. Line with 'combination' represents pairs of groups. The m/z-values and two statistical parameters, t-values and p-values, for all pairs of groups are arranged.</p> <pre> >group1 Sample1-1.mit Sample1-2.mit Sample1-3.mit Sample1-4.mit Sample1-5.mit >group2 Sample2-1.mit Sample2-2.mit Sample2-3.mit Sample2-4.mit Sample2-5.mit >group3 Sample3-1.mit Sample3-2.mit Sample3-3.mit Sample3-4.mit Sample3-5.mit >combination 1 vs 2 1 vs 3 2 vs 3 72.99098211 3.903 0.0022 3.817 0.0025 0.938 0.1878 73.65656016 1.236 0.1257 2.612 0.0154 1.892 0.0475 </pre>

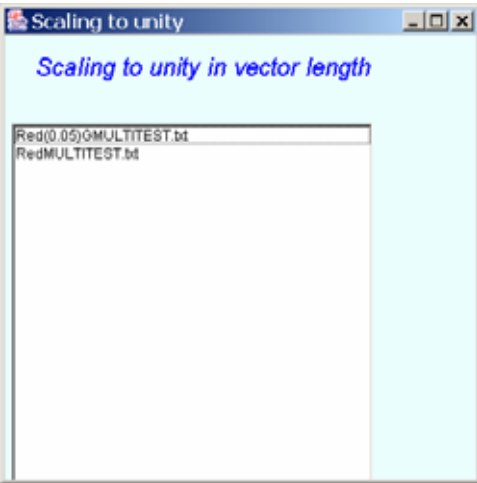
2.9 Peak Reduction

The m/z for statistically significant differences of the intensity between pairs of groups with P-value smaller than the threshold are selected. Thus, noisy intensities can be removed from the multivariate analysis.

Input file	PGMULTI, GMULTI
Output file	RED(thr) ; thr corresponds to P-value set by user.
Execution	<p>[1] Input threshold of P-value. [2] Click filename.</p> 
Output file format	(RED format described in 2.6 'M to R')

2.10. Scaling

For each m/z, peak intensities for multiple measurements are normalized to unity in sum of square.

Input file	RED
Output file	REDS
Execution	<p>[1] Click filename.</p> 
Output file format	(RED format described in 2.6 'M to R')

In M measurements, the intensities for individual m/z values are represented by a data matrix \mathbf{X} . Here, the number of IMCs are denoted by S, and the number of peaks is denoted by N.


$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ y_{N1} & \dots & y_{Nj} & \dots & y_{NM} \end{pmatrix} \quad (2.10.1)$$

For each m/z, peak intensities for multiple measurements are scaled by using Eq. (2.10.2).

$$x''_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^M x_{ij}}}$$


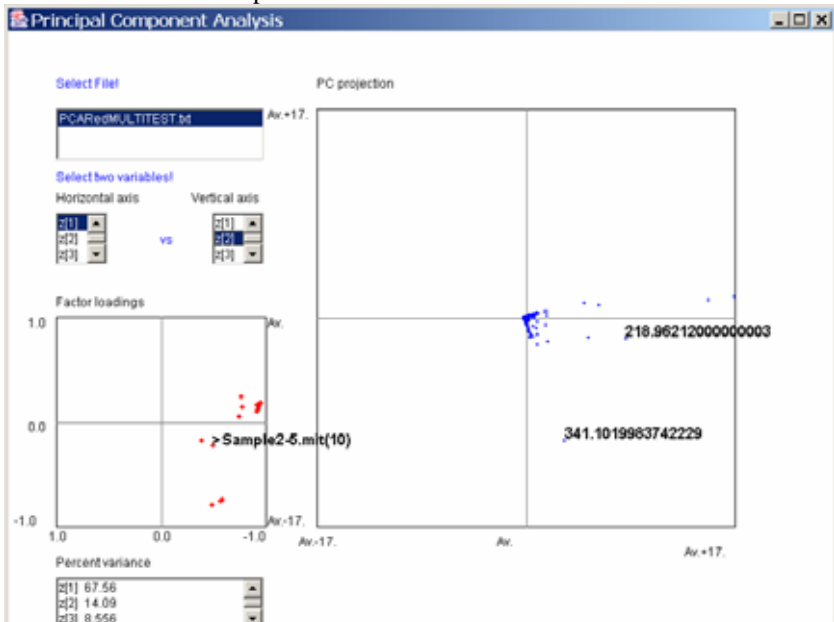
2.11. Pearson correlation

Pearson correlations of the intensities for pairs of m/z larger than the threshold set by the user are list up.

Input file	MULTI
Output file	Pearson(thr)
Execution	<p>[1] Input threshold of Correlation. [2] Click filename.</p> 
Output file format	<p>1st line started with '>Threshold' corresponds to the threshold set by the user. 2nd to the last lines correspond to pairs of peaks (m/z values) and their Pearson correlations. The numbers in parentheses represent index number for peaks in the input file. The index</p> <pre> >Threshold>=0.9 72.99283739615221(1) 109.48510305120904(7) 0.9802398346382768 72.99283739615221(1) 153.35659328616688(17) 0.9455749879863627 72.99283739615221(1) 166.05649834186937(19) 0.9856522752844742 </pre>

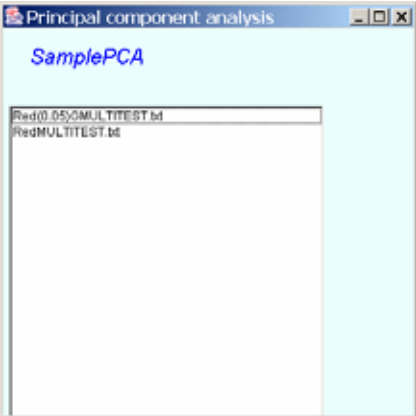
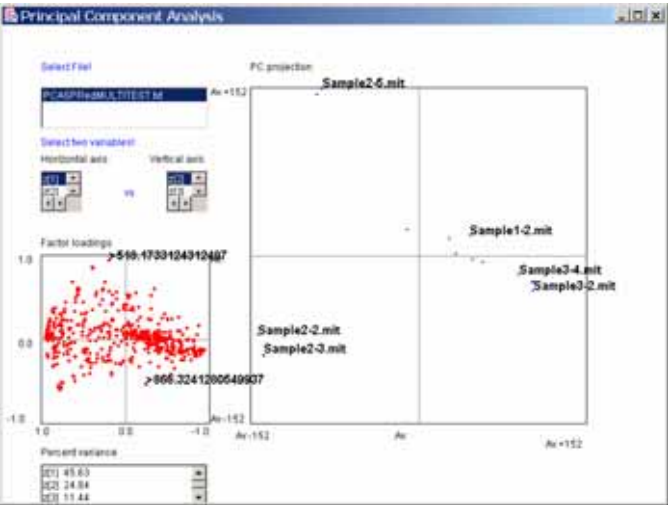
(iii) Multivariate Analysis**2.12. Peak-PCA and its Viewer**

Principal component analysis for peaks and visualization of its results. In PCA, three types of parameters, Score, Factor loading, and % Var are used for interpreting multivariate data. In Peak-PCA, variables correspond to experiments, and objects corresponds to peak intensities for individual m/z values. So Score is calculated for the peak intensities for each m/z, and Factor loading (correlation between s th principal component and t th variable) is calculated for pairs between a vector for s th principal components and a vector corresponding to t th experiment.

Input file	RED
Output file	PCA
Execution	<p>[1] Click filename.</p>  <p>[2] After finishing execution of Principal component analysis, the results of PCA can be visualized by clicking 'viewer' button on right side of 'Peak-PCA' button.</p> <p>[3] Select file and two variables, then we can obtain PC projection for m/z and Factor loadings for measurements. Information of m/z and measurements are obtained by clicking dots in the plots, PC projection and Factor loadings, respectively. Percent variance is also listed up in the left of the downside.</p> 

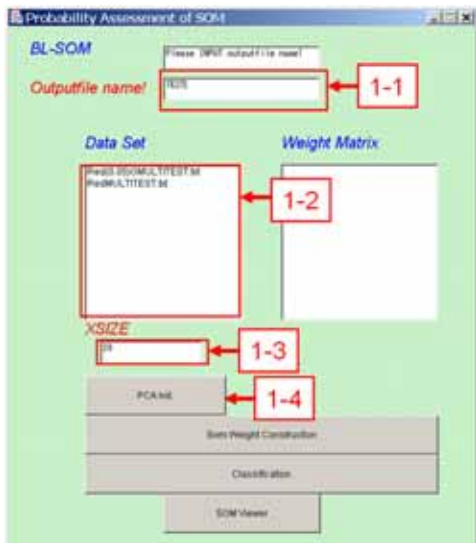
2.13. Sample-PCA and its Viewer

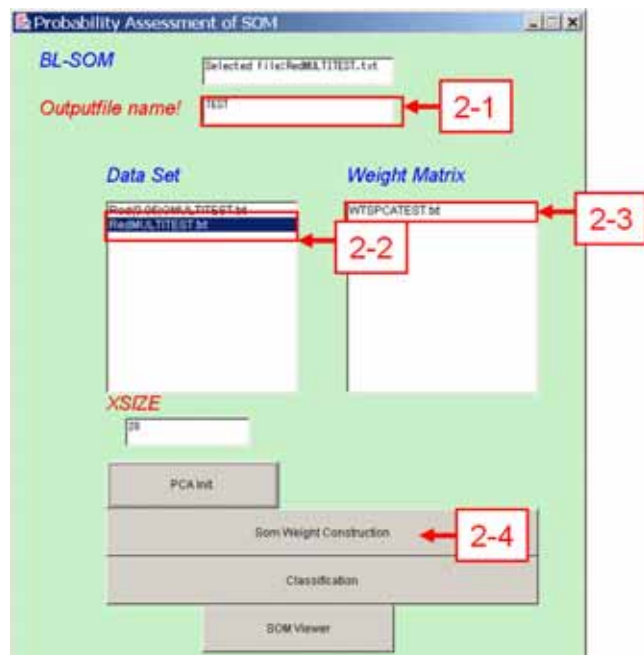
In Sample-PCA, variables correspond to m/z 's, and objects corresponds to peak intensities for individual experiments. So Score is calculated for the peak intensities for each experiment, and Factor loading is calculated for pairs between a vector for s th principal components and a vector corresponding to t th m/z .

Input file	RED								
Output file	PCASP								
Execution	<p>[1] Click filename.</p>  <p>[2] After finishing execution of Principal component analysis, the results of PCA can be visualized by clicking 'viewer' button on right side of 'Peak-PCA' button.</p> <p>[3] Select file and two variables, then we can obtain PC projection for measurements and Factor loadings for m/z. Information of measurements and m/z are obtained by clicking dots in the plots, PC projection and Factor loadings, respectively. Percent variance is also listed up in the left of the dbottomside.</p>  <table border="1" data-bbox="437 1727 560 1787"> <thead> <tr> <th colspan="2">Percent variance</th> </tr> </thead> <tbody> <tr> <td>(1)</td> <td>45.63</td> </tr> <tr> <td>(2)</td> <td>24.94</td> </tr> <tr> <td>(3)</td> <td>11.44</td> </tr> </tbody> </table>	Percent variance		(1)	45.63	(2)	24.94	(3)	11.44
Percent variance									
(1)	45.63								
(2)	24.94								
(3)	11.44								

2.14. BLSOM and its Viewer

Batch-learning SOM for peaks and visualization of its results

Input file	RED
Output file	CLSOM
Execution	<p>I. Construction of Self-organizing map</p> <p>1. Constructing of initial weight vectors by PCA</p> <p>[1-1] Input Outputfile name.</p> <p>[1-2] Select Inputdata</p> <p>[1-3] Set the number of weights in x size, then y size is automatically determined by variance ratio of the first and second principal components determined by PCA.</p> <p>[1-4] Click 'PCA Init.' button.</p>  <p>After execution of 'PCA Init.', an weight matrix file whose name is given by user and automatically added by 'WTSPCA' in the head is constructed in Weight Matrix.</p> <p>II Learning process by Data Set and Initial Weight Matrix</p> <p>[2-1] Input outputfile name.</p> <p>[2-2, 2-3] Select an inputfile and its corresponding initial weight matrix in Data Set and Weight Matrix, respectively.</p> <p>[2-4] Click 'Som Weight Construction' button.</p>



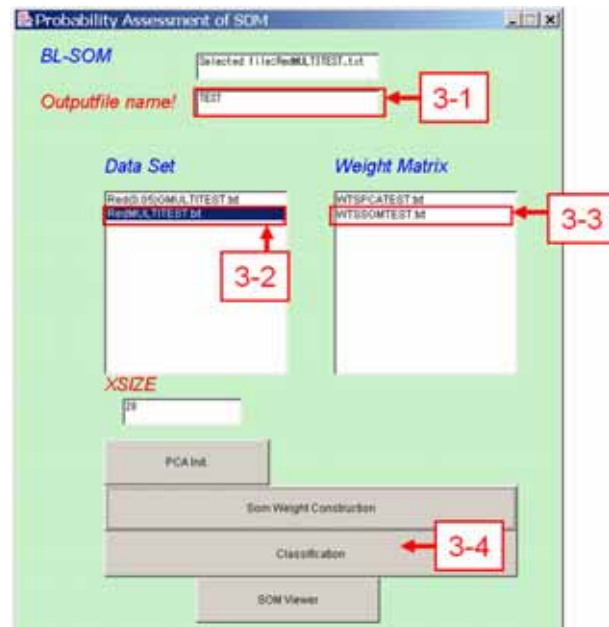
After execution, weight vectors optimized by input vectors are constructed in a filename with WTSSOM.

III Classification of objects (m/z)

[3-1] Input inputfile name.

[3-2, 3-3] Select an inputfile and its corresponding weight matrix started with WTSSOM in Data Set and Weight Matrix, respectively.

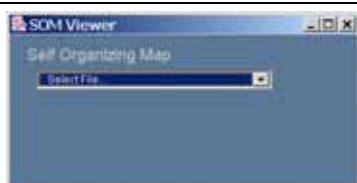
[3-4] Click 'Classification' button.



IV Visualization of Classification of objects (m/z)

(This process is the same as that of Viewer on the right of BL-SOM.

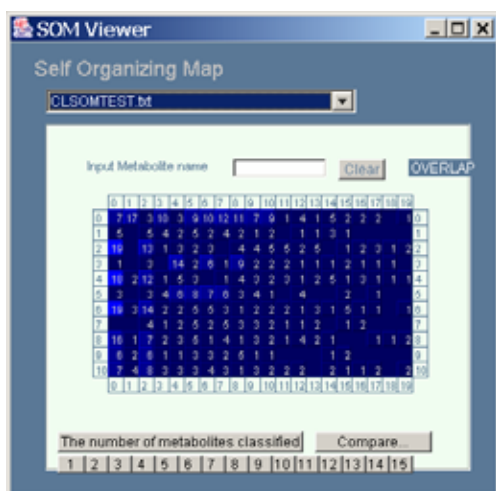
[4-1] Click 'SOM Viewer'.



[4-2] Select file.



Then, SOM Viewer is displayed.



Profile analysis

Click a square, m/z-values with similar profiles in multiple measurements are displayed. The following example is the profiles in the square at X=19 and Y=10. Two profiles in m/z with 255.2329 and 348.1023 are very similar in the multiple measurements.



Characterization of individual measurements

When a user wants to know high and low levels corresponding to individual experiments, he/she should click experiment ID. In this example the 5th experiment has been selected. Pink and Red lattices include only objects with measurements larger than the average for the selected experiment. Sky blue and Blue lattices include only objects with measurements smaller than the average for the selected experiment. A red lattice indicates that at least one of the objects belonging to it is with a measurement value larger than the average plus the standard deviation and a blue lattice indicates that at least one of

the objects belonging to it is with a measurement value smaller than the average minus the standard deviation.



BL-SOM package is also available in our laboratory <http://kanaya.aist-nara.ac.jp/SOM/>, and is applied to several works as bioinformatics tool as follows.

- 1 S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, T. Ikemura. Analysis of codon usage diversity for bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome., *Gene*, 276, 89-99 (2001)
- 2 T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, T. Ikemura, Informatics for unvailing hidden genome signature., *Genome Res.*, 13, 693-702 (2003).
- 3 M. Hirai, M. Yano, D. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, K. Saito, Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*, *Proc. Natl. Acad. Sci., USA*, 101, 10205-10210 (2004).
4. M. Hirai, M. Klein, Y. Fujikawa, M. Yano, D.B. Goodenowe, Y. Yamazaki, S. Kanaya, Y. Nakamura, M. Kitayama, H. Suzuki, N. Sakurai, D. Shibata, J. Tokuhisa, M. Reichelt, J. Gershenzon, J. Papenbrock, K. Saito, Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *J. Biol. Chem.*, 280,25590-5 (2005).