川原信夫	漢方薬に使用される生薬・薬用植物の現状	社団法人東京生薬協会会報	452 (1)	4-8	2012
Yoshimatsu, K.	Innovative cultivation: Hydroponics of medicinal plants in the closed-type cultivation facilities	*Journal of Traditional Medicines*	29	30-34	2012
Kawahara, N.	Project of Research Center for Medicinal Plant Resources, National Institute of Biomedical Innovation, Japan	*Tokyo Forum for International Standardization of Natural Medicines*		153-164	2012
川原信夫	東日本大震災後の国内薬用植物栽培の方向性を探る	薬用植物研究	34 (1)	1-3	2012
渕野裕之	独立行政法人医薬基盤研究所薬用植物資源研究センター筑波研究部栽培研究室－薬用植物成分研究と生薬の品質評価法の研究、種子の保存－	和漢薬	704 (1)	7-9	2012
熊谷健夫	医薬基盤研究所薬用植物資源研究センター筑波研究部栽培研究室－薬用植物の栽培研究と種子交換－	和漢薬	705 (2)	4-6	2012
菱田敦之	医薬基盤研究所薬用植物資源研究センター北海道研究部栽培研究室－日本における薬用植物栽培の普及とその課題－	和漢薬	706 (3)	4-7	2012
林 茂樹	薬用植物の品種育成について	和漢薬	707 (4)	4-6	2012
飯田 修	医薬基盤研究所薬用植物資源研究センター種子島研究部－熱帯、亜熱帯性薬用、有用植物の収集、保存、育成および利用	和漢薬	708 (5)	3-4	2012
杉村康司	薬用植物資源研究センター種子島研究部におけるソロモン諸島未利用植物資源の探索研究と絶滅危惧種タカクマムラサキの保存育成研究	和漢薬	709 (6)	6-9	2012
吉松嘉代他	植物工場での甘草生産に適したウラルカンゾウの選抜と育成	ブレインテクノニュース	149	1-9	2012
菱田敦之	アルテミシニンの生産を目的としたクソニンジンの栽培	道薬誌	29(2)	17-20	2012
菱田敦之	生薬「吉草根」の生産とその課題	道薬誌	29(4)	25-28	2012
菱田敦之	生薬「半夏」の生産とその課題	道薬誌	29(6)	23-26	2012
菱田敦之	薬用植物の栽培と今後の展望	農家の友	64 (12)	22-25	2012

吉松嘉代	甘草の水耕栽培 薬用植物資源の保護と確保	ファルマシア	49	141-146	2013
吉松嘉代	植物工場における薬用植物優良苗の育成と生産	*SHITA REPORT*	30	13-21	2013
林 茂樹	甘草の栽培について (前編)	道薬誌	30(2)	17-19	2013

J. Trad. Med. **27**, 210−216, 2010

Medical and
Pharmaceutical Society
for WAKAN-YAKU

Comparison of chemical constituents in *Glycyrrhiza uralensis* from various sources using a multivariate statistical approach

Ken Tanaka,[*a)] Atsutoshi Ina,[b)] Kosuke Hayashi,[a)] Katsuko Komatsu[a)]

[a)]*Division of Pharmacognosy, Department of Medicinal Resources, Institute of Natural Medicine, University of Toyama, 2630 Sugitani, Toyama 930-0194, Japan.* [b)]*Department of Research and Development, Kokando Co. Ltd., 2-9-1 Umezawacho, Toyama 930-0055, Japan.* *(Received July 22, 2010. Accepted October 21, 2010.)*

Abstract

Quality of Glycyrrhizae Radix samples collected in China and Mongolia was evaluated by multivariate statistical analysis of liquid chromatography-ion trap-time of flight (LC-IT-TOF) mass spectrometric data. In total, 17 peaks were annotated or identified in the chromatogram obtained from the analysis of a water extract of Glycyrrhizae Radix. The 10 compounds with the greatest degree of variance, (liquiritin apioside, liquiritin, glycyrrhizin, licoricesaponin A3, B2, E2, G2, H2, licorice glycoside E, and a compound having the same composition as glycyrrhizin) were selected as index compounds to create a data matrix for principal component analysis (PCA). Thirty three cultivated or wild Glycyrrhizae Radix samples, collected in China and Mongolia, were grouped characteristically by PCA. In addition, the possibility of the developing resources of Glycyrrhizae Radix in the eastern region of Mongolia is indicated as an alternative to Chinese Glycyrrhizae Radix.

Key words Glycyrrhizae Radix, multivariate statistical analysis, LC-IT-TOF MS, saponins.

Introduction

Glycyrrhizae Radix, an underground material derived from a species of the genus *Glycyrrhiza*, is one of the most popular crude drugs and has been used for over 2000 years in many Asian and European countries.[1)] Glycyrrhizae Radix shows a variety of pharmacological activities, and is applied to diseases of the stomach, liver, catarrh in respiratory organs and skin disorders in traditional Japanese medicines ("Kampo" medicines).[2)] These bioactivities are attributed to chemical constituents, such as triterpene saponins and falvonoids.[3)] It is well known that the level of active compounds varies widely depending on the plant species, geographic source, harvesting and processing.[4)] So far, several studies have been carried out on evaluating the quality of Glycyrrhizae Radix.[5)] In our previous paper, we have reported the differences of falvonoids and glycyrrhizin contents in Chinese and Mongolian Glycyrrhizae Radix.[6)] Kondo *et al.* have reported the constituent properties of Glycyrrhizae Radix derived from *G. uralensis, G. glabra* or *G. inflata* using the amounts of liquiritin, liquiritin apioside, liquiritigenin, isoliquiritin, isoliquiritin apioside, isoliquiritigenin, as well as glycyrrhizin, glycycoumarin, glabridin and licochalcon A as indices of the quality of the Radix.[7)] In addition, Wang *et al.* has reported on the development of the simultaneous quantification of liquiritin, liquiritigenin, isoliquiritigenin, glycyrrhizin, glycyrrhetinic acid and glycyrrhetinic acid methyl ester by HPLC.[8)] However, most of these studies have been focused on a small number of triterpene saponin and falvonoid in a methanol extract of Glycyrrhizae Radix and highly polar

*To whom correspondence should be addressed.
e-mail : ktanaka@inm.u-toyama.ac.jp

consists were not considered. Traditionally, crude drugs are used in a hot water decoction, and it is considered that the examination of the chemical constitutes in water extract is important to evaluate the quality of the crude drug as a medicine. Therefore, to reveal the comprehensive properties of crude drugs, the development of a no-targeting new evaluation strategy of water extract is still required.

Recently, LC-MS analysis, with principal component analysis (PCA) as the main statistical approach, aiming to identify and quantify all the components in the analytes, has aroused extensive interest in the field of botanical studies and it has been applied to the classification of plant material.[9-11] PCA is very useful for the detection of outliers and for finding patterns and trends. However, PCA can be considered to be the rotation of a data matrix to find the maximum variations in the observations, and the quality of PCA model is affected by the properties of the original data matrix, such as the variance of factors and correlation between the factors. Thus, careful selection of the constituent compounds and development of an informative data matrix are required to apply PCA to the classification of plant material.

In the present study, a chemometric approach, combining liquid chromatography-ion trap-time of flight mass spectrometric (LC-IT-TOF MS) analysis, was applied for discriminating the origin of Chinese and Mongolian Glycyrrhizae Radix.

Experimental

Materials and analytical sample preparation: Thirty three crude drug samples and plant specimens were purchased in markets or collected from the areas where they were grown (Table 1). The species of all samples and specimens were identified as *Glycyrrhiza uralensis* by morphological analysis and detection of glycycoumarin, a species specific compound, as shown in our previous paper.[6] All samples and specimens were deposited in the Museum of Materia Medica, Institute of Natural Medicine, University of Toyama (TMPW).

The samples were pulverized and the powder was screened through 150 μm sieves. Two grams of the fine powder was accurately weighed and extracted three times with 20 ml of water under reflux for 30 min. After centrifugation, the extracts were combined and lyophilized. The dried extract was dissolved in 70% of methanol at a concentration of 2 mg/ml and filtered through a 0.2 μm Millipore filter. One μl of this solution was injected into the LC-MS. The contents of glycyrrhizin, liquiritin and liquiritin apioside are indicated in our previous paper.[6]

Standard samples and reagents: Analytical grades of glycyrrhizin and liquiritin were purchased from Wako Chemical Co. Ltd (Osaka, Japan). Liquiritin apioside, licoricesaponin A3, B2, E2, G2, H2 and licorice glycoside E were isolated from a drug sample (TMPW No. 24238) purchased from Uchida Wakanyaku Co. Ltd. (Tokyo, Japan). The isolated compounds were identified by comparing their ^{1}H- and ^{13}C-NMR spectra with those reported in the literature.[12-15] Other analytical grade chemicals and HPLC grade chromatographic solvents were also purchased from Wako Chemical Co. Ltd (Osaka, Japan) or Nacalai Tesque, Inc. (Kyoto, Japan).

Analytical instruments: LC-MS analyses were performed with a Shimadzu LC-IT-TOF mass spectrometer, equipped with an ESI interface (Shimadzu, Kyoto, Japan). The ESI parameters were as follows: source voltage - 3.5 kV (negative mode), capillary temperature 200 ℃, nebulizer gas 1.5 l/min. The mass spectrometer was operated in negative ion mode scanning from *m/z* 100 to 1500. A Waters Atlantis T3 column (ODS column, 2.1 mm i.d. x 150 mm, Waters, MA) was used and the column temperature was maintained at 40 ℃. The mobile phase comprised a binary eluent of (A) 5 mM ammonium acetate solution, and (B) CH_3CN, under the following gradient conditions: 0-40 min linear gradient from 5 % to 100 % B, 40-60 min isocratic at 100 % B. The flow rate was 0.2 ml/min.

Chemometric data analysis: Thirty three samples were analyzed to identify potential discriminant variables. Peak finding in the total ion chromatogram, peak picking, alignment of the peak by the retention time and peak filtering of negative mass spectrometric raw data were carried out using Shimadzu LC-MS Solution software (Shimadzu, Kyoto, Japan). The detected peaks were filtered by two factors, retention time window

Table 1 Materials used in the present study

No.	Producing area	Purchased from	Collected date	TMPW[a] No.
1	Inner Mongolia, China (Cultivate)	Chifeng, Inner Mongolia, China	2002.9	21563
2	Inner Mongolia, China (Cultivate)	Huhehaote, Inner Mongolia, China	2002.9	21549
3	ibid.	ibid.	2002.9	21552
4	Inner Mongolia, China (Cultivate)	Chifeng, Inner Mongolia, China	2002.9	21585
5	Unknown	Taegu herbal drugs supermarket, Taegu, Korea	2003.8	22250
6	Inner Mongolia, China (Wild)	Uchida Wakanyaku Co. Ltd., Tokyo, Japan	2005.2	24113
7	ibid.	ibid.	2002.9	25883
8	ibid.	ibid.	2002.9	25884
9	Inner Mongolia, China (Wild)	Chifeng, Inner Mongolia, China	2002.9	21557
10	ibid.	ibid.	2002.9	21570-1
11	ibid.	ibid.	2002.9	21570-2
12	ibid.	ibid.	2002.9	21570-3
13	ibid.	ibid.	2002.9	21570-4
14	ibid.	ibid.	2002.9	21570-5
15	ibid.	ibid.	2002.9	21570-6
16	Ningxia, China (Wild)	Uchida Wakanyaku Co. Ltd., Tokyo, Japan	2004.11	23882
17	Gansu, China (Wild)	Tochimoto Tenkaido Co. Ltd., Tokyo, Japan	2004.11	23879-1
18	ibid.	ibid.	2004.11	23879-2
19	Inner Mongolia, China (Wild)	Uchida Wakanyaku Co. Ltd., Tokyo, Japan	2004.11	23881
20	Inner Mongolia, China (Wild)	Tochimoto Tenkaido Co. Ltd., Tokyo, Japan	2004.11	23880-1
21	ibid.	ibid.	2004.11	23880-2
22	Sergelen, Dornod Prov., Mongolia (Wild)		2004.7	M1209
23	Tmsagiyn hooly, Dornod Prov., Mongolia (Wild)		2004.7	M1218
24	Hyargas nuur, Uvs Prov., Mongolia (Wild)		2002.7	M795
25	ibid.		2002.7	M798
26	Sharga, Govi-Altay Prov., Mongolia (Wild)		2002.8	M863
27	ibid.		2002.8	M864
28	Buutsagaan-Bogd, Bayanhongor Prov., Mongolia (Wild))		2002.8	M891
29	Orog nuur, Bayanhongor Prov., Mongolia (Wild)		2002.8	M900
30	Bogd, Bayanhongor Prov., Mongolia (Wild)		2002.8	M904-1
31	ibid.		2002.8	M904-2
32	ibid.		2002.8	SN
33	Myangad, Ulaan hargana, Mongolia (Wild)		2003.10	22889

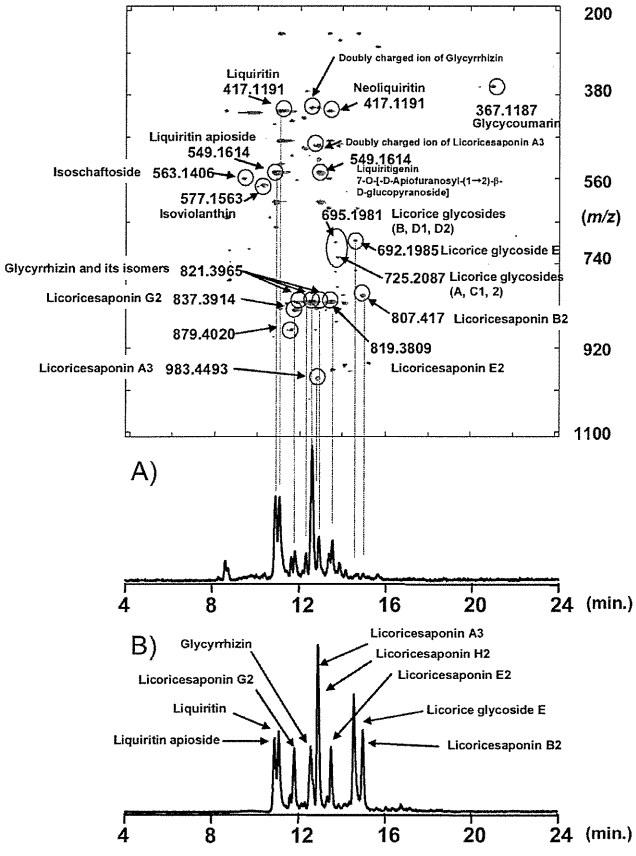
[a] The specimen reference number of the Museum of Materia Medica, Institute of Natural Medicine, University of Toyama (TMPW).

(retention time ± 10% of the pre-recorded reference peak) and similarity of the mass spectrum of the detected peak to the pre-recorded reference mass spectrum (a similarity index of more than 70%). The peak area values were converted to peak area values relative to the largest peak area amongst all 33 TIC. With this filtering, common peaks having equivalent quality in the TIC of the 33 samples were selected to construct an original data matrix. The variables in the original data matrix were aligned in the order of the extent of their variance and variables having small variance were removed to construct a data matrix for multivariate statistical analysis.

All the statistical analyses applied to the resulting 2D-data matrix, containing selected peaks (Table 2) and normalized intensities of the peaks, were carried out with Piroruet software (GL Science Inc, Tokyo).

Table 2 Compounds used as the factors in the principal component analysis and their mass spectral data.

	Retention time (min)	Compounds	Molecular formula	$(M-H)^-$ ions	Detected fragment ions in MS^2 from $(M-H)^-$
1	12.12	Glycyrrhizin	$C_{42}H_{62}O_{16}$	821.3965	803 $(M-H_2O-H)^-$, 759 $(M-H_2O-CO_2-H)^-$, 645 (M-dehydroxylatedGln-H)$^-$, 351 (di-glucuronic residue)
2	10.74	Liquiritin	$C_{21}H_{22}O_9$	417.1191	255 (liquiritigenin), 135 ($C_7H_3O_3^-$, $^{1,3}A^-$ fragment)
3	10.54	Liquiritin apioside	$C_{26}H_{30}O_{13}$	549.1614	429 ($^{0,2}X^-$ fragment), 417 (M-dehydoroxy apiose-H)$^-$, 297, 255 (liquiritigenin)
4	12.75	Licoricesaponin H2	$C_{42}H_{62}O_{16}$	821.3965	803 $(M-H_2O-H)^-$, 759 $(M-H_2O-CO_2-H)^-$, 645 (M-dehydroxy glucuronic acid-H)$^-$, 351 (di-glucuronic residue)
5	11.34	Licoricesaponin G2	$C_{42}H_{62}O_{17}$	837.3914	819 $(M-H_2O-H)^-$, 775 $(M-H_2O-CO_2-H)^-$, 661 (M-dehydroxylated glucuronic acid-H)$^-$, 351 (di-glucuronic residue)
6	12.39	Licoricesaponin A3	$C_{48}H_{72}O_{21}$	983.4493	821 (M-dehydroxy glucose-H)$^-$,803 $(M-H_2O-H)^-$, 759 $(M-H_2O-CO_2-H)^-$, 645 (M-dehydroxylated glucuronic acid-H)$^-$, 351 (di-glucuronic residue)
7	14.41	Licoricesaponin B2	$C_{42}H_{64}O_{15}$	807.4172	789 $(M-H_2O-H)^-$, 631 (M-dehydroxylated glucuronic acid-H)$^-$, 351 (di-glucuronic residue)
8	11.81	Isomer of glycyrrhizin	$C_{42}H_{62}O_{16}$	821.3965	803 $(M-H_2O-H)^-$, 759 $(M-H_2O-CO_2-H)^-$, 645 (M-dehydroxy glucuronic acid-H)$^-$, 351 (di-glucuronic residue)
9	13.05	Licoricesaponin E2	$C_{42}H_{60}O_{16}$	819.3809	801 $(M-H_2O-H)^-$, 757 $(M-H_2O-CO_2-H)^-$, 643 (M-dehydroxylatedGln-H)$^-$, 351 (di-glucuronic residue)
10	14.21	Licorice glycoside E	$C_{35}H_{35}NO_{14}$	692.1985	549 (M-indolecaboxy moiety)$^-$, 531 (M-indolecaboxylic acid -H)$^-$, 399 (531-dehydoroxy apiose)$^-$, 255(liquiritigenin)



Results and Discussion

Annotation of mass chromatographic fingerprint of Glycyrrhizae Radix water extract: For the identification of compounds in the water extract, a 2-D mass chromatographic fingerprint (TMPW no. 21557, sample 9 in Table 1) was created as shown in Figure 1. In this, the retention time of HPLC is indicated along the horizontal axis and *m/z* of the mass spectrometry is indicated along the vertical axis. The spots in the figure indicate the ions observed.

Annotation of the peaks was carried out by comparison of the HPLC retention time and mass spectra with those of the standard. In cases where comparison with

Figure 1 2-D Mass chromatographic fingerprints of the extract of Glycyrrhizae Radix (TMPW No.21557, sample 9 in Table 1) (A), total ion chromatogram of the extract of Glycyrrhizae Radix (TMPW No.21557, sample 9 in Table 1) (B) and total ion chromatogram of a mixture of glycyrrhizin, liquiritin, liquiritin apioside, licoricesaponin A3, B2, E2, G2, H2 and licorice glycoside E. High-resolution mass data and annotated compounds are indicated.

the standard was not available, annotation of the unknown compound was preferably assigned to a compound that had previously been reported from *G. uralensis* by comparing the high-resolution mass spectral data with those in the database.[16-18] In addition, the chemical structures were further confirmed by MS^n analysis.

Although a total of 17 peaks were annotated or identified as the major constituents in the mass chromatographic fingerprint, 10 peaks (liquiritin apioside, liquiritin, glycyrrhizin, licoricesaponin G2, A3, B2, E2, licorice glycoside E, and one compound having same composition as glycyrrhizin) were commonly observed through all 33 samples with large variances in their peak intensities (Table 2).

In Figure 2, total ion chromatograms and mass chromatograms monitored by the $(M-H)^-$ ion of glycyrrhizin,

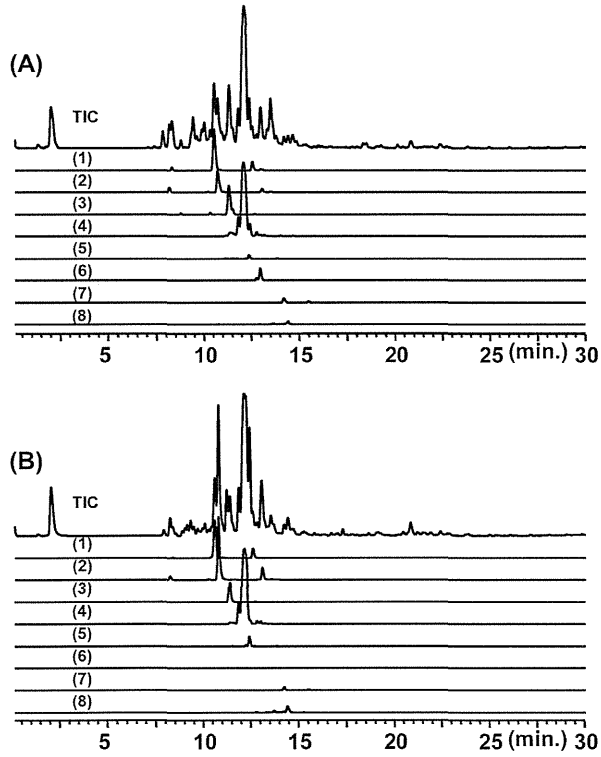


Figure 2 Total ion chromatograms of the extract of (A) Mongolian Glycyrrhizae Radix (TMPW M-904-1, sample 30 in Table 1) and (B) Chinese Glycyrrhizae Radix (TMPW No.23881, sample 19 in Table 1).
(1), *m/z* 549.1614 ($(M-H)^-$ of liquiritin apioside); (2), *m/z* 417.1191 ($(M-H)^-$ of liquiritin); (3), *m/z* 837.3914 ($(M-H)^-$ of licoricesaponin G2); (4), *m/z* 821.3965 ($(M-H)^-$ of glycyrrhizin and licoricesaponin H2); (5), *m/z* 983.4493 ($(M-H)^-$ of licoricesaponin A3); (6), *m/z* 819.3809 ($(M-H)^-$ of licoricesaponin E2); (7), *m/z* 692.1985 ($(M-H)^-$ of licorice glycoside E); (8), *m/z* 807.4172 ($(M-H)^-$ of licoricesaponin B2)

liquiritin, liquiritin apioside, licoricesaponin A3, B2, E2, G2, H2 and licorice glycoside E in extracts of Chinese and Mongolian samples are shown. Though small differences in the concentration of liquiritin and licoricesaponin A3 were observed, further visual differences between the chromatograms of the Chinese and Mongolian samples were not observed. Therefore, as an easier and non-biased interpretation of the results, and to reduce the dimensionality of the multivariate data obtained from the LC-MS results, multivariate statistical analysis, PCA, was used.

Principal component analysis: PCA transforms a number of correlated variables into a smaller number of uncorrelated synthesis variables known as principal components, which are linear combinations of the correlated variables. The first principal component accounts for the greatest extent of the variability in the data. The subsequent combinations describe the maximum amount of remaining variability. PCA, as with any other multivariate statistical method, is sensitive to missing data, and poor linear correlation between variables, due to poorly distributed variables. As a result, the quality of the original data matrix has a large impact upon PCA.

The process used in this study was as follows:
(1) Peak picking from the total collection of TIC and creation of a data matrix
Using the mass spectral similarity of the peaks, the 17 peaks commonly observed in the TIC of all 33 samples were selected to construct a row data matrix, the columns of which represent the variables (peak area), whilst the rows contain the samples included for analysis. The variables were aligned in the order of the extent of their variance. Ten variables, having larger variance, were selected to reduce the dimension of the row data.
(2) PCA on the new data set
The data matrix, $\boldsymbol{X}$, was decomposed into the product of a score matrix, $\boldsymbol{T}$, and a loading matrix, $\boldsymbol{L}$, by PCA. The columns of $\boldsymbol{L}$ are the principal components (PCs), the new factors of which are linear combinations of the original variables. The score matrix, $\boldsymbol{T}$, is the projection of the samples onto the axes defined by the loadings. Each sample has a coordinate on each new axis. The score plots are used to reveal similarities among

samples by their measured properties. The distribution of the samples on this graph reveals a pattern that correlates to the general characteristics of the samples.

In Figure 3, the PCA scores plot is shown, where each Glycyrrhizae Radix sample is represented by a marker. The first three PCs account for 73.8 % of the total variance (PC1, 38.2 %; PC2, 19.7 %; PC3, 15.9 %). The scores plot clearly indicates that Chinese (triangle and square markers in Figure 3) and Mongolian (circle markers in Figure 3) Glycyrrhizae Radix have different properties. Furthermore, it is clear that the cultivated (square markers in Figure 3) and wild (triangle markers in Figure 3) Glycyrrhizae Radix in the Chinese samples can apparently be further classified. In the PCA analysis, the peaks with large loading values can be considered to be markers that strongly contribute to the classification of the samples. In Figure 4, the loading values for PC1 and PC2 are shown. From the analysis of the score and loading results, the following differences in the properties of the Glycyrrhizae Radix derived from different regions was observed:

(1) The cultivated and wild samples are classified according to their position on the PC1 axis. The amounts of glycyrrhizin, the compound with the same composition as glycyrrhizin, liquiritin apioside, liquiritin, licoricesaponin A3, E2, G2 and licorice glycoside E contribute to PC1. Chinese cultivated Glycyrrhizae Radix contains lower amounts of these compounds than wild Glycyrrhizae Radix.

(2) The Chinese and Mongolian samples are classified according to their position on the PC2 axis. Chinese samples contain larger amounts of liquiritin, licoricesaponin A3 and E2. On the other hand, Mongolian samples contain larger amounts of licoricesaponin H2 and licorice glycoside E.

It is interesting that two Mongolian samples grown in the eastern region of Mongolia (samples 22 and 23) show similar constituent patterns to Chinese wild Glycyrrhizae Radix. This result suggested that Glycyrrhizae Radix from the east of Mongolia could be an alternative candidate to Chinese Glycyrrhizae Radix.

In conclusion, we examined a multivariate statistical approach for evaluating Glycyrrhizae Radix obtained from different locations using LC-IT-TOF MS. Ten compounds with large variance (liquiritin apioside, liquiritin, glycyrrhizin, a compound having the same composition as glycyrrhizin, licoricesaponin A3, B2, E2, G2, H2 and licorice glycoside E) were selected as index compounds to create a data matrix for PCA. Thirty three cultivated or wild Glycyrrhizae Radix samples collected in China and Mongolia were classified according to their characteristics. It was clarified that

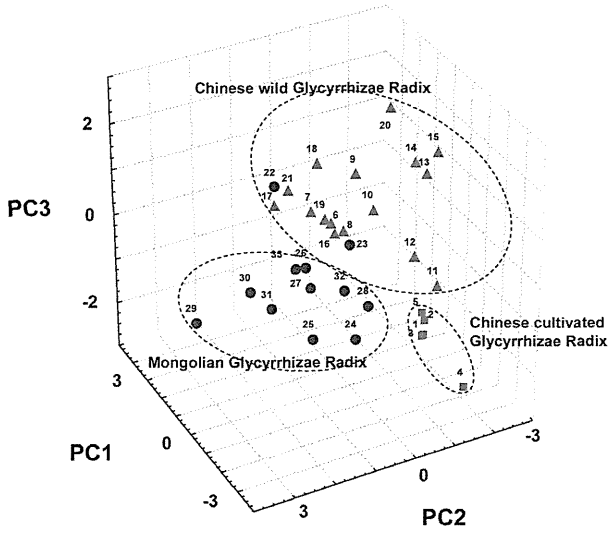


Figure 3 PCA scores plot for the Chinese and Mongolian Glycyrrhizae Radix samples
▲, Chinese wild Glycyrrhizae Radix; ■, Chinese cultivated Glycyrrhizae Radix; ●, Mongolian Glycyrrhizae Radix; The numbers in the scores plot indicate the sample No. in Table 1.

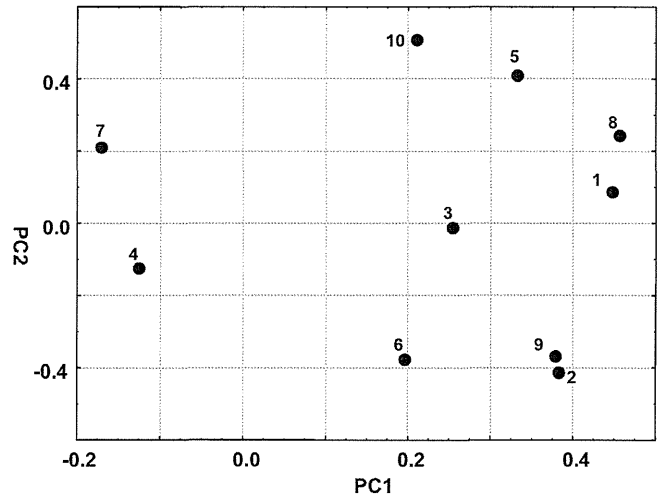


Figure 4 PCA loading plot of the Chinese and Mongolian Glycyrrhizae Radix samples
The numbers in the loading plot indicate the compound No. in Table 2.

Chinese cultivated and wild Glycyrrhizae Radix can be discriminated by the amounts of glycyrrhizin, liquiritin apioside, liquiritin, licoricesaponin A3, E2, G2 and licorice glycoside E and that Chinese and Mongolian Glycyrrhizae Radix are distinguishable by the amounts of liquiritin, licoricesaponin A3, E2 and H2 and licorice glycoside E. Furthermore, the possibility is indicated of the development of resources of Glycyrrhizae Radix in the eastern region of Mongolia as an alternative to Chinese Glycyrrhizae Radix.

References

1) Zhang, Q. and Ye, M.: Chemical analysis of the Chinese herbal medicine Gan-Cao (licorice). *J. Chromatogr.* A., **1216**, 1954-1969, 2009.
2) Society of Japanese Pharmacopoeia, The Japanese Pharmacopoeia, 15th ed., The Ministry of Health, Labour and Welfare (English edition), Tokyo, p. 1026, 2007.
3) Asl, M. N. and Hosseinzadeh, H.: Review of pharmacological effects of Glycyrrhiza sp. and its bioactive compounds. *Phytother. Res.,* **22**, 709-724, 2008.
4) Liang, Y. Z., Xie, P. and Chan, K.: Quality control of herbal medicines. *J. Chromatogr.,* B. **812**, 53-70, 2004.
5) Kitagawa, I., Chen, W., Taniyama, T., Harada, E., Hori, K., Kobayashi, M. and Ren, J.: Quantitative determination of constituents in various licorice roots by means of high performance liquid chromatography. *Yakugaku Zasshi,* **118**, 519-528, 1998.
6) Zhu, S., Sugiyama, R., Batkhuu, J., Sanchir, C., Zou, K., Komatsu, K.: Survey of Glycyrrhizae Radix resources in Mongolia: chemical assessment of the underground part of Glycyrrhiza uralensis and comparison with Chinese Glycyrrhizea Radix. *J. Nat. Med.,* **63**, 137-146, 2009.
7) Kondo, K., Shiba, M., Nakamura, R., Morota, T. and Shoyama, Y.: Constituent properties of licorices derived from *Glycyrrhiza uralensis, G. glabra,* or *G. inflata* identified by genetic information. *Biol. Pharm. Bull.,* **30**, 1271-1277, 2007.
8) Wang, Y. and Yang, Y.: Simultaneous quantification of flavonoids and triterpenoids in licorice using HPLC. *J. Chromatogr.* B., **850**, 392-399, 2007.
9) Lavine, B. and Workman, J.: Chemometrics. *Anal. Chem.,* **80**, 4519-4531, 2008.
10) Woo, Y. A., Kim, H. J., Cho, J. H. and Chung, H.: Discrimination of herbal medicines according to geographical origin with near infrared reflectance spectroscopy and pattern recognition techniques. *J. Pharm. Biomed. Anal.,* **21**, 407-413, 1999.
11) Yang, J., Chen, L. H., Zhang, Q., Lai, M. X. and Wang, Q.: Quality assessment of cortex cinnamomi by HPLC chemical fingerprint, principle component analysis and cluster analysis. *J. Sep. Sci.,* **30**, 1276-1283, 2007.
12) Yahara, S. and Nishioka, I.: Flavonoid glucosides from licorice. *Phytochemistry,* **23**, 2108-2109, 1984.
13) Kitagawa, I., Hori, K., Sakagami, M., Zhou, J. and Yoshikawa, M.: Saponin and sapogenol. XLVIII. On the constituents of the roots of *Glycyrrhiza uralensis* Fischer from northeastern China. (2). Licorice-saponins D3, E2, F3, G2, H2, J2, and K2. *Chem. Pharm. Bull.,* **41**, 1337-1345, 1993.
14) Kitagawa, I., Hori, K., Sakagami, M., Hashiuchi, F., Yoshikawa, M. and Ren, J.: Saponin and sapogenol. XLIX. On the constituents of the roots of *Glycyrrhiza inflata* Batalin from Xinjiang, China. Characterization of two sweet oleanane-type triterpene oligoglycosides, apioglycyrrhizin and araboglycyrrhizin. *Chem. Pharm. Bull.,* **41**, 1350-1357, 1993.
15) Hatano, T., Takagi, M., Ito, H. and Yoshida, T.: Acylated falvonoid glycosides and accompanying phenolics from licorice. *Phytochemistry,* **47**, 287-293, 1998.
16) Buckingham, J.: The Dictionary of Natural Products. Chapman & Hall/CRC, Florida, 2007.
17) Nakamura, Y., Asahi, H., Amin, M. A., Kurokawa, K. and Kanaya, S.: A Comprehensive Species-Metabolite Relationship Database (KNApSAcK), http://kanaya.naist.jp/ KNApSAcK/.
18) Meng, X., Li, H., Song, F., Liu, C., Liu, Z. and Liu, S.: Studies on Triterpenoids and Flavones in *Glycyrrhiza uralensis* Fisch. by HPLC-ESI-MSn and FT-ICR-MSn. *Chinese J. Chem.,* **27**, 299-305, 2009.

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).

Journal of MASS SPECTROMETRY

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,

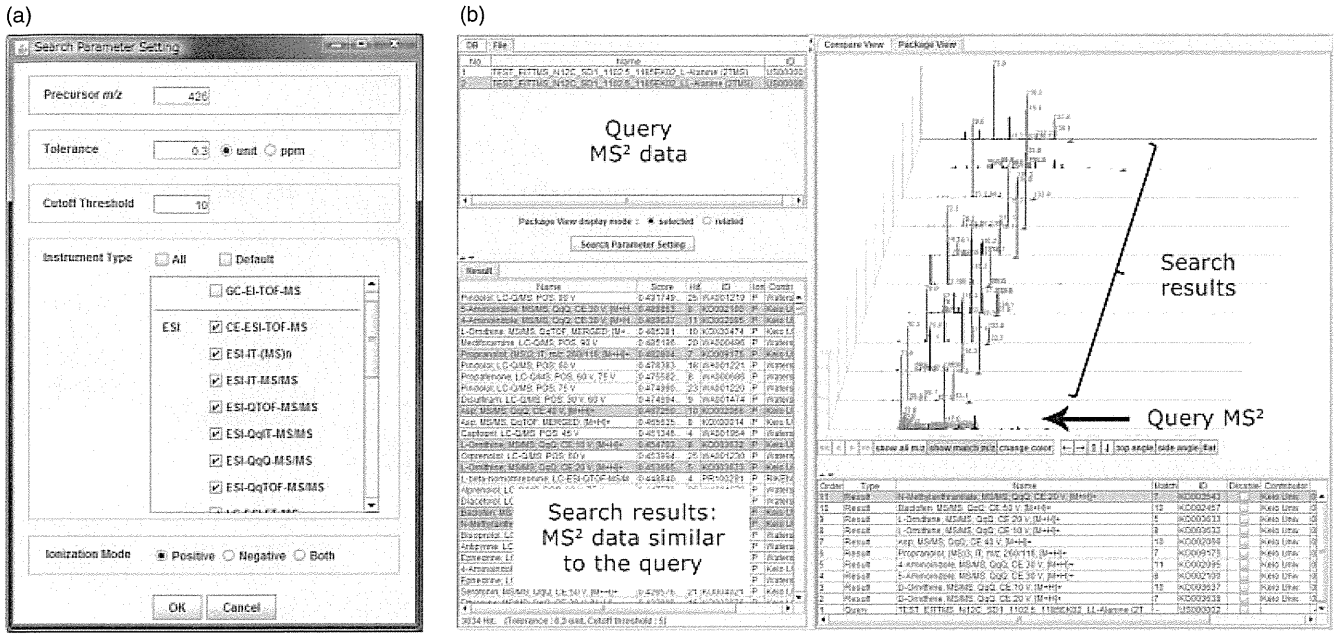


Figure 2. Search Parameter Setting and Spectral Search. (a) Search parameters are selected and input on the applet. The 'Precursor ion' is specified by the *m/z* value. 'Tolerance' is the error allowance of *m/z* values. When a peak in the query data and the corresponding peak in the target data have 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.

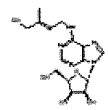
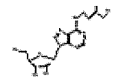
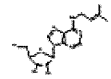
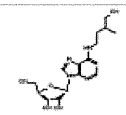
	Name	Formula / Structure	ExactMass	ID
☐	⊟ **trans-Zeatin-riboside** 1 spectrum	C15H21N5O5 	351.15427	
☐	LC-ESI-QTOF-MS/MS; CE:Ramp 5-60 V; [M+H]+			PR100209
☐	⊟ **trans-Zeatin riboside** 4 spectra	C15H21N5O5 	351.15427	
☐	LC-MS/MS; QqQ; CE:40.0 eV; [M+H]+			PR020095
☐	LC-MS/MS; QqQ; CE:30.0 eV; [M+H]+			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

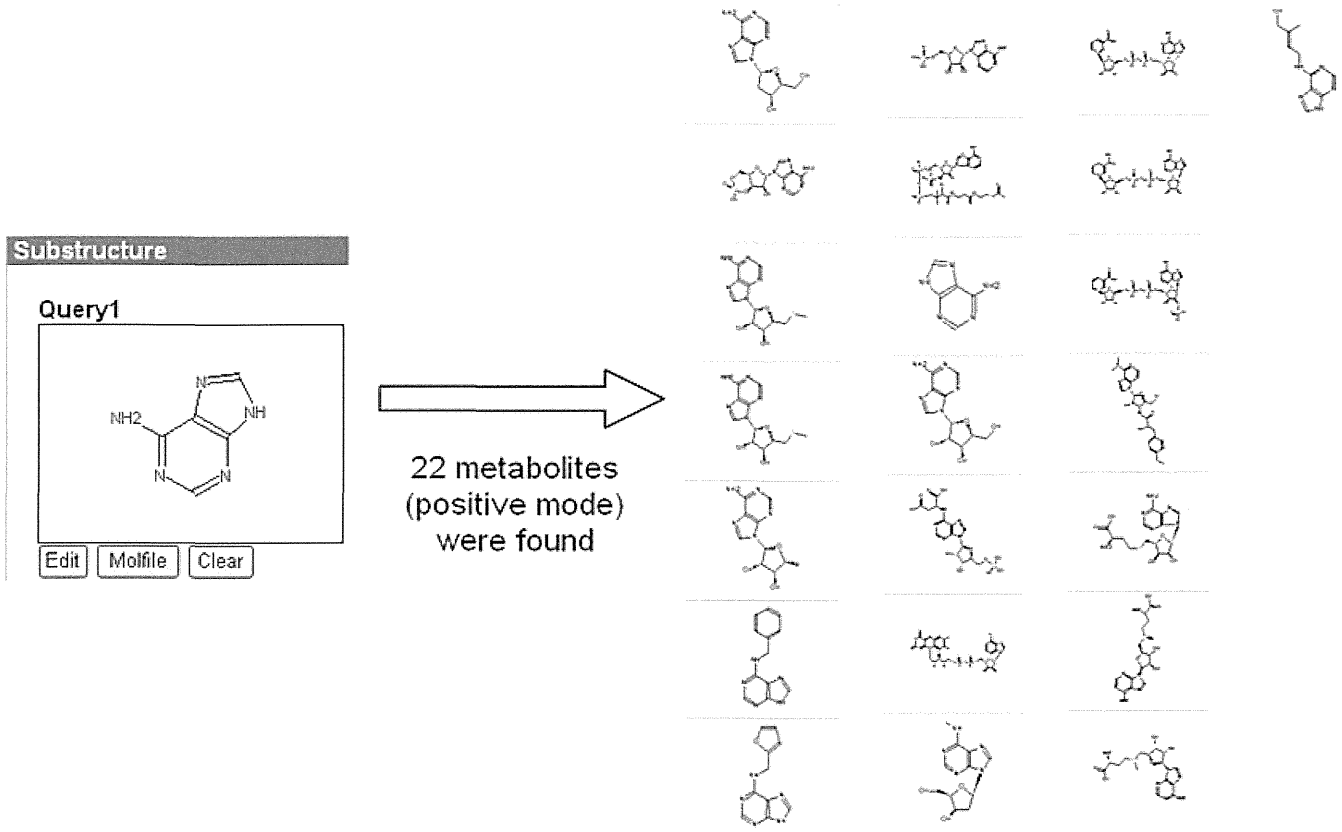


Figure 4. Substructure Search. When a substructure is submitted as a query, all chemical structures containing the query substructure are listed.

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

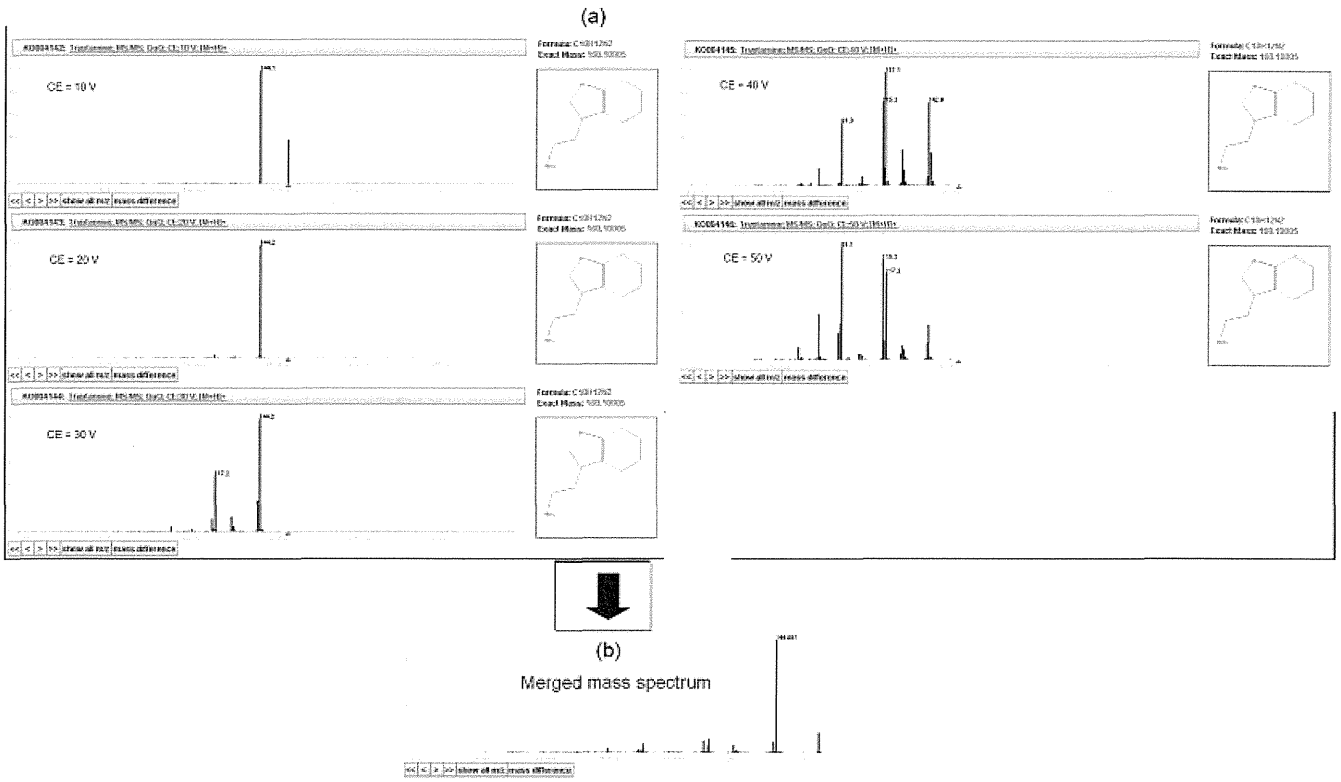


Figure 5. Merged mass spectral data. (a) The mass spectra of tryptamine analyzed on ESI-QqQ-MS^2 by different collision energies (CE), 10–50 V. (b) The five ESI-QqQ-MS^2 data of tryptamine were overlaid and merged into one 'merged mass spectrum'.

as the mislabeling of a test tube are caught by comparing the observed mass of the molecular ions with the calculated mass from the molfiles. For higher resolution MS^n data of known chemical compounds, chemical formulae may be uniquely assignable to most of the product ions in a higher *m/z* range within an error range of 50 ppm. Contributors are advised to add the chemical annotation of as many product ions as possible in an optional data field, PK$ANNOTATION, of the MassBank record format. Such chemical annotations are useful for the removal of mass spectral data that contain ions from contaminants. Annotations are also helpful to MassBank contributors who evaluate the mass accuracy of the data.

At present, MassBank data are the mass spectra of specific chemical compounds commercially available as purified reagents of metabolites. In the near future, we will accept the mass spectral data of metabolites detected and identified by LC-MS^n analysis of biological cell and tissue samples. In such cases, contributors must provide satisfactory experimental evidence for the identification of the chemical compounds in the chemical section.[11] We will also accept LC-, GC- and CE-coupled or direct MS^n data analysis of tissue pieces or single cells. Such data will include the mass spectra of identified and unidentified chemical compounds. Identified chemical compounds are indicated by their chemical names or structures and unidentified or unknown chemical compounds by their MS^n data. MS^n data are used as the tag of unidentified chemical compounds. By comparing the MS^n data analyzed on different biological samples, the intersample similarity or difference of the chemical compounds can be determined.

Sharing mass spectral data among research communities

Beginning in June 2008, the Mass Spectrometry Society of Japan supported MassBank as the official database of the society. In the near future, the society's journal will recommend the authors to register their mass spectral data in MassBank at the time they submit their manuscripts. MassBank will provide the authors with accession numbers for citation of the data in the manuscript. This will make it possible for readers to lookup data details on MassBank and to search for related articles with Spectral Search and other search tools available on MassBank. We plan to advocate the registration of mass spectral data in MassBank among contributors to other academic journals. In 2009 we started collaboration with LipidBank, the official database of the JCBL and organized joint special lectures on MassBank and LipidBank at annual meetings. The society and the conference will work jointly to seek continuous academic funding to support both MassBank and LipidBank.

MassBank provides a record field for copyright, the default holders of which are contributors, but none for data distribution. Because the distribution of mass spectral data is another method of data sharing, users and contributors propose to prepare a record field for data distribution in which contributors express under the terms of the Creative Commons Attribution Licenses.[34] We will prepare the record field and an FTP site for the download of data. Additionally, we consider augmenting the record documentation of MassBank by conforming to the guidelines for the controlled vocabularies from Proteomics Standards Initiative (PSI).[35]

Conclusions

MassBank is based on the three concepts. First, it is a public database of mass spectral data analyzed under nonstandardized experimental conditions. Second, it is a distributed database in which contributors prepare and provide their data from their own data servers on the Internet. Third, it develops and provides free tools for contributors to prepare and manage data on their sites. To improve the metabolite identification from mass spectra, we merged ESI-MS^2 data of identical chemical compounds analyzed under different experimental conditions. Merged data as a TS of spectral search were significantly improved precision without decreasing recall of the spectral search when compared with the unmerged original data set. This showed that merging spectral data is useful for generating reference data for metabolite identification.

Acknowledgements

We thank Ms Rie Matsuzawa and Ms Michi Kittaka for their input of data, Dr Toshiaki Katayama for his advice on the development of Web API on MassBank, Dr Zenzaburo Tozuka and Dr Yoshinao Wada for their support on introducing MassBank as the official database of the Mass Spectrometry Society of Japan and Dr Masaru Tomita for his financial support. This work was supported by a grant for the advancement and standardization of biological databases (2006-2010) from the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency (to T. N., M. A. and S. K.), a grant-in-aid for Scientific Research on the Priority Area from the Ministry of Education, Culture, Sports, Science and Technology of Japan (grant number 18016028 to T. N., M. A. and S. K.), research grants from Yamagata Prefecture and Tsuruoka City (to T. N. and M. A.) and a grant from the New Energy and Industrial Technology Development Organization (NEDO) of Japan as part of the 'Development of Fundamental Technologies for Controlling the Material Production Process of Plants' (to T. A., N. S., H. S. and D. S.).

References

[1] P. N. Schofield, T. Bubela, T. Weaver, L. Portilla, S. D. Brown, J. M. Hancock, D. Einhorn, G. Tocchini-Valentini, M. Hrabe de Angelis, N. Rosenthal. Post-publication sharing of data and tools. *Nature* **2009**, *461*, 171.

[2] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, D. Steinhauser. GMD@CSB.DB: the Golm Metabolome database. *Bioinformatics* **2005**, *21*, 1635.

[3] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, G. Siuzdak. METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **2005**, *27*, 747.

[4] A. Kameyama, N. Kikuchi, S. Nakaya, H. Ito, T. Sato, T. Shikanai, Y. Takahashi, K. Takahashi, H. Narimatsu. A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. *Anal. Chem.* **2005**, *77*, 4719.

[5] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, I. Forsythe. HMDB: a knowledge base for the human metabolome. *Nucleic Acids Res* **2009**, *37*, D603.

[6] National Institute of Standards and Technology, NIST Standard Reference Database 1A, NIST/EPA/NIH Mass Spectral Library with Search Program: (Data Version: NIST 08, Software Version 2.0f). http://www.nist.gov/srd/nist1a.htm. [Last accessed: March 2010].

[7] National Institute of Advanced Industrial Science and Technology, Japan. Spectral Database for Organic Compounds, SDBS. http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi. [Last accessed: March 2010].

[8] C. M. Dobson. Chemical space and biology. *Nature* **2004**, *432*, 824.

[9] J. Clardy, C. Walsh. Lessons from natural molecules. *Nature* **2004**, *432*, 829.

[10] S. A. Sansone, T. Fan, R. Goodacre, J. L. Griffin, N. W. Hardy, R. Kaddurah-Daouk, B. S. Kristal, J. Lindon, P. Mendes, N. Morrison, B. Nikolau, D. Robertson, L. W. Sumner, C. Taylor, M. van der Werf, B. van Ommen, O. Fiehn. The metabolomics standards initiative. *Nat. Biotechnol.*. **2007**, *25*, 846.

[11] L. Sumner, A. Amberg, D. Barrett, M. Beale, R. Beger, C. Daykin, T. Fan, O. Fiehn, R. Goodacre, J. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. Lane, J. Lindon, P. Marriott, A. Nicholls, M. Reily, J. Thaden, M. Viant. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3*, 211.

[12] F. Matsuda, K. Yonekura-Sakakibara, R. Niida, T. Kuromori, K. Shinozaki, K. Saito. MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. *Plant J.* **2009**, *57*, 555.

[13] A. Sreekumar, M. Poisson, T. M. Rajendiran, A. P. Khan, Q. Cao, J. Yu, B. Laxman, R. Mehra, R. J. Lonigro, Y. Li, M. K. Nyati, A. Ahsan, S. Kalyana-Sundaram, B. Han, X. Cao, J. Byun, G. S. Omenn, D. Ghosh, S. Pennathur, D. C. Alexander, A. Berger, J. R. Shuster, J. T. Wei, S. Varambally, C. Beecher, A. M. Chinnaiyan. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **2009**, *457*, 910.

[14] L. A. McDonnell, R. M. A. Heeren. Imaging mass spectrometry. *Mass Spectrom. Rev.* **2007**, *26*, 606.

[15] T. Masujima. Live single-cell mass spectrometry. *Anal. Sci.* **2009**, *25*, 953.

[16] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **2010**, *38*, D355.

[17] United States National Library of Medicine, National Institutes of Health, National Center for Biotechnology Information. PubChem Compounds Database. http://pubchem.ncbi.nlm.nih.gov/. [Last accessed: March 2010].

[18] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, S. Kanaya. KNApSAcK: A comprehensive species-metabolite relationship database. In *Plant Metabolomics*, K. Saito, R. A. Dixon, L. Willmitzer(Eds). Springer-Verlag Berlin: NY, **2006**, 165.

[19] Japanese Conference on the Biochemistry of Lipids. Database of natural lipids. http://www.lipidbank.jp/. [Last accessed: March 2010].

[20] E. Fahy, S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. van Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, E. A. Dennis. A comprehensive classification system for lipids. *J. Lipid Res.* **2005**, *46*, 839.

[21] D. J. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.

[22] International Union of Pure and Applied Chemistry. The IUPAC International Chemical Identifier. http://www.iupac.org/inchi/. [Last accessed: March 2010].

[23] S. Tanaka, K. Aoshima, Y. Miura, Y. Oda. 57th ASMS Conference on Mass Spectrometry and Allied Topics (American Society for Mass Spectrometry), Philadelphia, PA, 31 May to 04 June, **2009**.

[24] Biomarkers and Personalized Medicine Core Function Unit, Eisai Product Creation Systems, Eisai Co. Ltd. Mass++. http://groups.google.com/group/massplusplus. [Last accessed: March 2010].

[25] Institute for Advanced Biosciences, Keio University. MassBank. http://www.massbank.jp. [Last accessed: March 2010].

[26] Leibniz Institute of Plant Biochemistry, Stress and Developmental Biology. MassBank. http://msbi.ipb-halle.de/MassBank/. [Last accessed: March 2010].

[27] S. E. Stein, D. R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859.

[28] H. Horai, M. Arita, T. Nishioka. Comparison of ESI-MS in MassBank Database. 1st International Conference on BioMedical Engineering and Informatics, Sanya, Hainan, China, 28–30

